STATISTICAL MULTIRESOLUTION DANTZIG ESTIMATION IN IMAGING: FUNDAMENTAL CONCEPTS AND ALGORITHMIC FRAMEWORK

KLAUS FRICK

Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen

PHILIPP MARNITZ

Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen

AXEL MUNK

Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen
and

Max Planck Institute for Biophysical Chemistry
Am Faßberg 11, 37077 Göttingen

ABSTRACT. In this paper we are concerned with fully automatic and locally adaptive estimation of functions in a "signal + noise"-model where the regression function may additionally be blurred by a linear operator, e.g. by a convolution. To this end, we introduce a general class of *statistical multiresolution estimators* and develop an algorithmic framework for computing those. By this we mean estimators that are defined as solutions of convex optimization problems with ℓ_∞-type constraints. We employ a combination of the alternating direction method of multipliers with Dykstra's algorithm for computing orthogonal projections onto intersections of convex sets and prove numerical convergence. The capability of the proposed method is illustrated by various examples from imaging and signal detection.

1. Introduction

In numerous applications, the relation of observable data Y and the (unknown) signal of interest u^0 can be modeled as an inverse linear regression problem. We shall assume that the data $Y = \{Y_{\boldsymbol{\nu}}\}$ is sampled on the equidistant grid $X = \{1, \ldots, m\}^d$, with $m, d \in \mathbb{N}$ and that $u^0 \in U$ for some linear space U, such as the Euclidean space or a Sobolev class of functions. Hence the model can be formalized as

$$Y_{\boldsymbol{\nu}} = (Ku^0)_{\boldsymbol{\nu}} + \varepsilon_{\boldsymbol{\nu}}, \quad \boldsymbol{\nu} \in X. \tag{1}$$

Here we assume that $\varepsilon = \{\varepsilon_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu} \in X}$ are independent and identically distributed r.v. with $\mathbf{E}(\varepsilon_{\boldsymbol{\nu}}) = 0$ and $\mathbf{E}(\varepsilon_{\boldsymbol{\nu}}^2) = \sigma^2 > 0$ (white noise). Moreover, $K : U \to (\mathbb{R}^m)^d$ denotes a linear operator that encodes the functional relation between the quantities that are accessible by experiment and the underlying signal. Often the operator K does not have a continuous inverse (or its inverse is ill-conditioned in a discrete setting, where K is a matrix), that is estimation of u^0 given the data Y is an *ill-posed problem*. As a consequence, estimators for u^0 can not be obtained by merely applying the inverse of K to an estimator of Ku^0, in general. Instead, more sophisticated *statistical regularization* techniques have to be employed that, loosely speaking, are capable of simultaneously inverting K and solving the regression problem.

The application we primarily have in mind is the reconstruction of low-dimensional signals (e.g. images) u^0 which are presumed to exhibit a strong neighborhood structure as it is characteristic for imaging or signal detection problems. These neighborhood relations are often modeled by prior smoothness or structural assumptions on u^0 (e.g. on the texture of an image).

The aim of this paper is twofold. First, we will introduce the broad class of *statistical multiresolution estimators (SMRE)*. We claim that numerous regularization techniques, that were recently proposed for different problems in various branches of applied mathematics and statistics, can be considered as special cases of these. Among others, this includes the *Dantzig selector* (see [4, 7, 29] and references therein) that was recently proposed in the context of high dimensional statistics. Our prior focus, however, will be put on imaging problems and it will turn out that the aforementioned neighborhood relations can be modeled within our SMRE framework in a straightforward manner. This will result in *locally adaptive* and *fully automatic* image reconstruction methods.

The high intrinsic structure of the signals that are typically under consideration in imaging is in contrast to the usual situation in high-dimensional statistics. Here u^0 is usually assumed to be unstructured but to have a sparse representation with respect to some basis of U (cf. [7, 8, 43]). Consequently, the consistent estimation of u^0 is realized by minimizing a regularization functional which fosters sparsity, such as the ℓ_1-norm of the coefficients, subject to an ℓ_∞-constraint on the coefficients of the residual, i.e.

$$\inf_{u \in U} \|u\|_1 \quad \text{s.t.} \quad \|K^*(Y - Ku)\|_\infty \leq q. \tag{2}$$

In order to apply this approach for image reconstruction, two modifications become necessary: Often one aims to minimize other regularization functionals such as the total variation semi-norm (cf. [34, 37]) or Sobolev norms, say. Hence, we suggest to replace the ℓ_1-norm in (2) by a general convex functional J that models the smoothness or texture information of signals or images (cf. [1, 35]). Furthermore, we relax the ℓ_∞-constraint such that neighborhood relations of the image can be taken into account. This generalizes the Dantzig selector to

this task in a natural way and obviously increases estimation efficientcy. As we will layout in Paragraph 1.2.2, this requires new algorithms to compute efficiently the resulting large scale optimization problem.

1.1. **Statistical Multiresolution Estimation.** We will now introduce the announced class of estimators. To this end, let $\mathcal{S}$ be some index set and $\mathcal{W} = \left\{ \omega^S \ : \ S \in \mathcal{S} \right\}$ be a set of given weight-functions on the grid $X = \{1, \ldots, m\}^d$. A *statistical multiresolution estimator (SMRE)* (or *generalized Dantzig selector*), is defined as a solution of the constrained optimization problem

$$\inf_{u \in U} J(u) \quad \text{s.t.} \quad \max_{S \in \mathcal{S}} \left| \sum_{\boldsymbol{\nu} \in X} \omega^S_{\boldsymbol{\nu}} \left(\Lambda(Y - Ku) \right)_{\boldsymbol{\nu}} \right| \leq q. \tag{3}$$

Here, $J : U \to \mathbb{R}$ denotes a regularization functional that incorporates a priori knowledge on the unknown signal u^0 (such as smoothness) and $\Lambda : (\mathbb{R}^m)^d \to (\mathbb{R}^m)^d$ a possibly non-linear transformation. The constant q can be considered as a universal *regularization parameter* that governs the trade-off between regularity and data-fit of the reconstruction. In most practical situations q is chosen to be the α-quantile q_α of the *multiresolution (MR) statistic* $T(\varepsilon)$, where $T : (\mathbb{R}^m)^d \to \mathbb{R}$ encodes the inequality constraint in (3), i.e.

$$T(v) = \max_{S \in \mathcal{S}} \left| \sum_{\boldsymbol{\nu} \in X} \omega^S_{\boldsymbol{\nu}} \left(\Lambda(v) \right)_{\boldsymbol{\nu}} \right|, \quad v \in (\mathbb{R}^m)^d. \tag{4}$$

To this end, we assume the distribution of $T(\varepsilon)$ to be (approximately) known. This can either be obtained by simulations or in some cases the limiting distribution can even be derived explicitly. The regularization parameter q then admits a sound statistical interpretation: Each solution $\hat{u}_\alpha$ of (3) satisfies

$$\mathbb{P} \left(J(\hat{u}_\alpha) \leq J(u^0) \right) \geq \alpha$$

i.e. the estimator $\hat{u}_\alpha$ is *more regular* (in terms of J) than u^0 with a probability of at least α. To see this simply observe that the true signal u^0 satisfies the constraint in (3) with probability at least α.

For a given estimator $\hat{u}$ of u^0, the set $\mathcal{W}$ is assumed to be rich enough in order to catch all relevant non-random signals that are visible in the residual $Y - K\hat{u}$. Then, the average function

$$\mu_S(v) = \left| \sum_{\boldsymbol{\nu} \in X} \omega^S_{\boldsymbol{\nu}} \left(\Lambda(v) \right)_{\boldsymbol{\nu}} \right| \tag{5}$$

evaluated at $v = Y - K\hat{u}$ is supposed to be significantly larger than q for at least one $\omega \in \mathcal{W}$, whenever $Y - K\hat{u}$ fails to resemble white noise. Put differently, the MR-statistic $T(Y - K\hat{u})$ is bounded by q, whenever $Y - K\hat{u}$ is accepted as white noise according to the *resolution* provided by $\mathcal{W}$. In fact, this is a key observations that reveals numerous potential application areas of the estimation method (3). The examples we have in mind are mainly from *statistical signal detection and imaging*, where the index set $\mathcal{S}$ is typically chosen to be an overlapping (redundant) system of subsets of the grid X and ω^S is the normalized indicator function on $S \in \mathcal{S}$. Consequently the inequality constraint in (3) guarantees that the residual resembles white noise on all sets $S \in \mathcal{S}$. In other words, the SMRE approach in (3) yields a reconstruction method that *locally adapts the amount of regularization* according to the underlying image features. We illustrate this in Section 3 by various examples.

Summarizing, the optimization problem in (3) amounts to choose the most parsimonious among all estimators $\hat{u}$ for which the residual $Y - K\hat{u}$ resembles white noise according to the statistic T. If $Y - K\hat{u}$ contains some non randon signal, i.e. $T(Y - K\hat{u})$ is likely to be larger than q and u happens to lie outside the admissible domain of (3). Thus, the multi-resolution constraint prevents too parsimonious reconstructions due to the minimization of J.

1.2. **Algorithmic Challenges and Related Work.**

1.2.1. *Multiresolution Methods.* SMREs and related MR statistics have recently been studied in various contexts. We give a brief (but incomplete) overview.

Classical MR statistics are obtained from the general form in (4) by setting $U = (\mathbb{R}^m)^d$ and $\Lambda = \mathrm{Id}$. Moreover, one considers the system $\mathcal{W}$ to contain indicator functions on cubes. To be more precise, define the index set $\mathcal{S}$ to be the system of all d-dimensional cubes in X and set $\omega^S = \chi_S / \sqrt{\#S}$. Then, the MR-statistic in (4) reduces to

$$T(v) = \max_{S \in \mathcal{S}} \frac{1}{\sqrt{\#S}} \left| \sum_{\boldsymbol{\nu} \in S} v_{\boldsymbol{\nu}} \right|.$$

This statistic was introduced in [42] (called scanning statistic there) in order to detect a signal against a noisy background. It was shown in [31] that

$$\lim_{m \to \infty} \frac{T(\varepsilon)}{\sqrt{2d \log m}} = \sigma \quad \text{a.s.}$$

If the system $\mathcal{S}$ is reduced to the set of all *dyadic* squares, then it was proved in [28] that (after suitable transformations) T also converges weakly to the Gumbel distribution. There, the authors also established a method for locally adaptive image denoising employing linear diffusion equations with spatially varying diffusivity. SMREs (3) have been studied recently for the case $d = 1$ in [12] and [6], where total-variation penalty and the number of jumps in piecewise constant regression were considered as regularization functional J, respectively. In [21] consistency and convergence rates for SMREs have been studied in a general Hilbert space setting.

SMREs with squared residuals, that is $\Lambda(v)_{\boldsymbol{\nu}} = v_{\boldsymbol{\nu}}^2$, yield another class of estimators that have attracted much attention. Above all, the situation where $\mathcal{S}$ consists of the single set X and ω^X is chosen to be the constant 1 function is of special interest, since then (3) reduces to the *penalized least square estimation*. In particular (3) then can be rewritten into

$$\inf_{u \in U} J(u) + \lambda \sum_{\boldsymbol{\nu} \in X} (Ku - Y)_{\boldsymbol{\nu}}^2 \tag{6}$$

for a suitable multiplier $\lambda > 0$. If $J(u) = \|u\|_1$ the LASSO estimator will result (cf. [43]). Recently, also non-trivial choices of $\mathcal{S}$ were considered. In [3] $\mathcal{S}$ is chosen to consist of a partition of G which is obtained beforehand by a Mumford-Shah segmentation. In [15], a subset $S \subset X$ is fixed and afterwards $\mathcal{S}$ is defined as the collection of all translates of S.

In [17] MR-statistics are used for shape-constrained estimation based on testing qualitative hypothesis in nonparametric regression for $d = 1$. Here, the weight functions ω^S incorporate qualitative features such as monotonicity or concavity. Similarly, MR-statistics are used in [18] in order to detect locations of local increase and decrease in density estimation. Much in the same spirit is the work in [16] where multiscale sign tests are employed for computing confidence bands for isotonic median curves.

As mentioned previously, the *Dantzig-selector* [7] is also covered by the general SMRE framework in (3). To see this, set $U = \mathbb{R}^p$ (with typically $p \gg m$), $\Lambda = \mathrm{Id}$ and define the weights

$$\omega^S = K\chi_S, \quad S \in \mathcal{S}.$$

Then, each solution of (3) can be considered as a generalized Dantzig selector. The matrix $K \in \mathbb{R}^{m \times p}$ in this context is usually interpreted as *design matrix* of a high dimensional linear model. The classical Dantzig selector as introduced in [7] then results in the special case where $\mathcal{S}$ only consists of single-elemented subsets of $\{1, \ldots, p\}$ and J is chosen to be the ℓ_1-regularization functional

$$J(u) = \|u\|_1 = \sum_{i=1}^{p} |u_i|.$$

Hence LASSO and Dantzig selector are uni-scale estimators which take into account the largest ($\mathcal{S} = \{X\}$) and smallest ($\mathcal{S}$ consists of all singletons in $\{1, \ldots, p\}$) scales, respectively. In this sense, they constitute two extreme cases of SMRE.

1.2.2. *Algorithmic Challenges.* From a computational point of view, computing an SMRE amounts to solve the *constrained optimization problem* (3) which can be rewritten into

$$\inf_{u \in U} J(u) \quad \text{s.t.} \quad \mu_S(Y - Ku) \leq q, \ \forall (S \in \mathcal{S}). \tag{7}$$

We note that in practical applications the number of constraints in (7), that is the cardinality of the index set $\mathcal{S}$, can be quite large (in Section 3.2 denoising of a 512×512 image results in more than 6 million inequalities). Moreover, the inequalities (even for the simplest case where $\Lambda = \mathrm{Id}$) are mutually correlated. Both of these facts turn (7) into a numerically challenging problem and standard approaches (such as interior point or conjugate gradient methods) perform far from satisfactorily.

The authors in [3, 15, 28] approach the numerical solution of (7) by means of an analogon of (6) with spatially dependent multiplier $\lambda \in (\mathbb{R}^m)^d$, i.e.

$$\inf_{u \in U} J(u) + \sum_{\boldsymbol{\nu} \in X} \lambda_{\boldsymbol{\nu}} (Ku - Y)_{\boldsymbol{\nu}}^2.$$

Starting from a (constant) initial parameter $\lambda = \lambda_0$, the parameter λ is iteratively adjusted by increasing it in regions which were poorly reconstructed before according to the MR-statistic T. This approach strongly depends on the special structure of $\mathcal{S}$ that allows a straightforward identification of each set $S \in \mathcal{S}$ with a unique point in the grid X. Put differently, it is not clear how to modify this paradigm in order to solve (3) for highly redundant systems $\mathcal{S}$ as we have it in mind.

Recently a general algorithmic framework was introduced in [2] for the solutions of large-scale convex cone problems

$$\inf_{u \in U} J(u) \quad \text{s.t.} \quad Ku - Y \in \mathcal{K}$$

where $\mathcal{K}$ is a convex cone in some Euclidean space. The approach was realized in the software package *Templates for First-Order Conic Solvers (TFOCS)* [1]. The above formulation is very

---

[1]available at http://tfocs.stanford.edu/

general and in order to recover (7) one has to consider the cone

$$\mathcal{K} = \left\{ (v, q) \in (\mathbb{R}^m)^d \times \mathbb{R} \ : \ \left| \sum_{\boldsymbol{\nu} \in S} \Lambda(v)_{\boldsymbol{\nu}} \right| \leq q \ \forall (S \in \mathcal{S}) \right\}$$

The approach in [2] employs the dual formulation of the problem

$$\inf_{\xi \in V} J^*(K^* v) + \langle Y, v \rangle \quad \text{s.t. } v \in \mathcal{K}^*$$

which involves the computation of the dual cone $\mathcal{K}^*$ (J^* denotes the Legendre-Fenchel dual of J). This approach is particularly appealing for the uni-scale Dantzig selector since in this situation the cone $\mathcal{K}$ coincides with the epi-graph of the ℓ^∞-norm and hence its dual cone is straightforward to compute (it is the epi-graph of the ℓ^1-norm). As it is argued in [2], this approach is capable of computing Dantzig selectors for large scale problems in contrast to previous approaches such as standard linear programming techniques [7] or homotopy methods such as DASSO [29] or [41]. As the authors stress, their approach works well in the case when $\mathcal{K}$ is the epi-graph of a norm for which the projections onto $\mathcal{K}^*$ are tractable and computationally efficient. However, for the applications we have in mind (such as locally adaptive imaging reconstruction), the approach in [2] is only of limited use: In contrast to the aforementioned epi-graphs, the large number of (strongly dependent) constraints in (7) brings about a cone $\mathcal{K}$ that on the one hand exhibits a tremendous amount of faces compared to the dimension of the image space $\dim(H) = md$ and that on the other hand is no longer symmetric w.r.t. to the q-axis. Both of these facts turn the computation of dual cone $\mathcal{K}^*$ (or the projections onto it) into a most challenging problem, even in the simplest case when Λ is linear.

The aim of this paper is to develop a general algorithmic framework that makes solutions of (7) numerically accessible for many applications. In order to do so we propose to introduce a slack variable in (7) and then use the *alternating direction method of multipliers*, an Uzawa-type algorithm that decomposes problem (7) into a J-penalized least squares problem for the primal variable and a orthogonal projection problem on the feasible set of (7) for the slack variable. This approach has the appealing effect that once an implementation for the projection problem is established, different regularization functionals J can easily be employed without changing the backbone of the algorithm. Our work is much in the same spirit as [33], which considered an alternating direction method for the computation of the Dantzig selector recently. In this case the computation of the occurring orthogonal projections are available in closed form, whereas in our applications this is not the case due to the aforementioned dependencies.

In order to tackle the orthogonal projection problem we employ Dykstra's projection method [5] which is capable of computing the projection onto the intersection of convex bodies by merely using the individual projections onto the latter. The efficiency of the proposed method hence increases considerably if the index set $\mathcal{S}$ can be decomposed into "few" partitions that contain indices of mutually independent inequalities in (7). In particular, by this approach we will be able to compute classical SMRE (as introduced in [12, 21]) in $d = 2$ space dimensions which to our knowledge has never been done so far. This puts us into the position to study the performance of such estimators compared with other benchmark methods in locally adaptive signal recovery (such as *adaptive weights smoothing* cf. [40]). As it will turn out in Section 3 it will outperform these visually as well as quantitatively.

1.3. **Organization of the Paper.** The paper is organized as follows: In Section 2 we introduce a general algorithmic approach for computing SMREs. We will rewrite (7) into a linearly constrained problem and compute a saddle point of the corresponding augmented Lagrangian by the alternating direction method of multipliers in Paragraph 2.2. Under quite general assumption, we prove convergence of the algorithm in Theorem 2.2 and give some qualitative estimates for the iterates in Theorem 2.4. One of the occurring minimization steps amounts to the computation of an orthogonal projection onto a convex set in Euclidean space. In Paragraph 2.3, this problem will be tackled by means of Dykstra's projection algorithm introduced in [5]. Finally, we illustrate the performance of some particular instances of SMREs in Section 3: we study problems in nonparametric regression, image denoising and deconvolution of fluorescence microscopy images and compare our results to other methods by means of simulations.

2. Computational Methodology

In this section we will address the question on how to solve the linearly constrained optimization problem (7). After discussing some notations and basic assumptions in Subsection 2.1, we will reformulate the problem in Paragraph 2.2 such that the alternating direction method of multipliers (ADMM), a Uzawa-type algorithm, can be employed as a solution method. As an effect, the task of computing a solution of (7) is replaced by alternating

   i) solving an unconstrained penalized least squares problem that is *independent of the MR-statistic T* and

  ii) computing the orthogonal projection on a convex set in Euclidean space that is *independent of J*.

This reveals an appealing modular nature of our approach: The regularization functional J can easily be replaced once a method for the projection problem is settled. For the latter we will propose an iterative projection algorithm in Paragraph 2.3 that was introduced by Boyle and Dykstra in [5].

2.1. **Basic Assumptions and Notation.** From now on, X will stand for the d-dimensional grid $\{1, \ldots, m\}^d$ and agree upon $H = \mathbb{R}^X \simeq (\mathbb{R}^m)^d$ being the space of all real valued functions $v : X \to \mathbb{R}$. Moreover, we assume that $\mathcal{S}$ denotes some index set and that $\mathcal{W} = \{\omega^S \; : \; S \in \mathcal{S}\}$ is a collection of elements in H. For two elements $v, w \in H$ we will use the standard inner product and norm

$$\langle v, w \rangle = \sum_{\boldsymbol{\nu} \in X} v_{\boldsymbol{\nu}} w_{\boldsymbol{\nu}} \quad \text{and} \quad \|v\| = \sqrt{\langle v, v \rangle}$$

respectively. Next, we assume that $\Lambda : H \to H$ is continuous such that $\Lambda(0) = 0$ and that for all $S \in \mathcal{S}$ the mapping

$$v \mapsto \langle \omega^S, \Lambda(v) \rangle$$

is convex. With this notation, we can rewrite the average function in (5) in the compact form

$$\mu_S(v) = \left| \langle w^S, \Lambda(v) \rangle \right|.$$

We note, that it is not restricitve to consider more generaly $\Lambda : H \to \mathbb{R}^N$ with arbitrary $N \in \mathbb{N}$. This could e.g. be useful for augmenting the constraint set of (7) with further constraints of different type. For the signal and image detection problems as studied in this paper, however, Λ is always a pointwise transformation of the residuals. Hence, we will restrict our considerations on the case when $\Lambda : H \to H$.

Furthermore, we define U to be a separable Hilbert-space with inner product $\langle \cdot, \cdot \rangle_U$ and induced norm $\| \cdot \|_U$. The operator $K : U \to H$ is assumed to be linear and bounded and the functional $J : U \to \mathbb{R}$ is convex and lower semi-continuous, that is

$$\{u_n\}_{n \in \mathbb{N}} \subset U \text{ and } \lim_{n \to \infty} u_n =: u \in U \quad \Longrightarrow \quad J(u) \leq \liminf_{n \to \infty} J(u_n).$$

Recall the definition of the MR-statistic in (4). Throughout this paper we will agree upon the following

Assumption A. *i) For all $y \in H$ there exists $u \in U$ such that $T(Ku - y) < q$.*
ii) For all $y \in H$ and $c \in \mathbb{R}$ the set

$$\left\{ u \in U \; : \; \max_{S \in \mathcal{S}} \mu_S(Ku - y) + J(u) \leq c \right\}$$

is bounded.

Under Assumption A it follows from standard techniques in convex optimization, that a solution of (7) exists. As we will discuss in Section 2.2 it even follows that a saddle point of the corresponding Lagrangian exists (cf. Theorem 2.1 below). In this context Assumption A i) is often referred to as *Slater's constraint qualification* and is for instance satisfied if $K(U)$ is dense in H. Moreover, Assumption A ii) will be needed in order to guarantee convergence of the algorithm for computing such a solution, as it is proposed in the upcoming section. This requirement is fulfilled if J is coercive i.e.

$$\lim_{\|u\|_U \to \infty} J(u) = \infty.$$

In many applications U is some function space and J a gradient based regularization method, such as the total variation semi-norm (cf. Section 3.2). Then a typical sufficient condition for Assumption A ii) is that K does not annihilate constant functions.

2.2. **Alternating Direction Method of Multipliers.** By introducing a slack variable $v \in H$ we rewrite (7) to the equivalent problem

$$\inf_{u \in U, v \in H} J(u) + G(v) \quad \text{subject to} \quad Ku + v = Y. \tag{8}$$

Here, G denotes the characteristic function on the feasible region $\mathcal{C}$ of (7), that is,

$$\mathcal{C} = \{v \in H \; : \; \mu_S(v) \leq q \; \forall (S \in \mathcal{S})\} \quad \text{and} \quad G(v) = \begin{cases} 0 & \text{if } v \in \mathcal{C} \\ +\infty & \text{else.} \end{cases} \tag{9}$$

Note that due to the assumptions on Λ, the set $\mathcal{C}$ is closed and convex. The technique of rewriting (7) into (8) is referred to as the *decomposition-coordination approach*, see e.g. Fortin & Glowinski [20, Chap. III]. There, Lagrangian multiplier methods are used for solving (8). To this end, we recall the definition of the *augmented Lagrangian* of Problem (8), that is

$$L_\lambda(u, v; p) = \frac{1}{2\lambda} \|Ku + v - Y\|^2 + J(u) + G(v) - \langle p, Ku + v - Y \rangle, \quad \lambda > 0. \tag{10}$$

The name stems from the fact that the ordinary Lagrangian

$$L(u, v; p) = J(u) + G(v) - \langle p, Ku + v - Y \rangle$$

is augmented by the quadratic penalty term $(2\lambda)^{-1} \|Ku + v - Y\|^2$ that fosters the fulfillment of the linear constraints in (8). The *augmented Lagrangian method* consists in computing a saddle point $(\hat{u}, \hat{v}, \hat{p})$ of L_λ, that is

$$L_\lambda(\hat{u}, \hat{v}; p) \leq L_\lambda(\hat{u}, \hat{v}; \hat{p}) \leq L_\lambda(u, v; \hat{p}), \quad \forall ((u, v, p) \in U \times H \times H)$$

We note that each saddle point $(\hat{u}, \hat{v}, \hat{p})$ of the augmented Lagrangian L_λ is already a saddle point of L and vice versa and that in either case the pair $(\hat{u}, \hat{v})$ is a solution of (8) (and thus $\hat{u}$ is a desired solution of (7)). This follows e.g. from [20, Chap 3. Thm. 2.1]. Sufficient conditions for the existence of saddle points are usually harder to come up with. Assumption A summarizes a standard set of such conditions.

Theorem 2.1. *Assume that Assumption A holds. Then, there exists a saddle point $(\hat{u}, \hat{v}, \hat{p})$ of L_λ.*

Proof. According to [19, Chap. III, Prop. 3.1 and Prop. 4.2] a saddle point of L exists, if there is an element $u_0 \in U$ such that G is continuous at $Ku_0 - Y$ and that

$$\lim_{\|u\|_Q \to \infty} J(u) + G(Ku - Y) = \infty. \tag{11}$$

According to Assumption A i) and due to the continuity of Λ the first requirement is clearly satisfied. Further, the coercivity assumption (11) is a consequence of Assumption A ii). $\square$

We will use the *Alternating Direction Method of Multipliers (ADMM)* (cf. Algorithm 1) as proposed in [20, Chap. III Sec. 3.2] for the computation of a saddle point of L_λ (and hence of a solution of (7)): Successive minimization of the augmented Lagrangian L_λ w.r.t. the first and second variable followed by an explicit step for maximizing w.r.t. the third variable is performed. Convergence of this method is established in Theorem 2.2 which is a generalization of [20, Chap. III Thm. 4.1]. We note that the proof, as presented in the Appendix A allows for *approximate* solution of the individual subproblems. For the sake of simplicity, we present the Algorithm in its exact form.

Theorem 2.2. *Every sequence $\{(u_k, v_k)\}_{k \geq 1}$ that is generated by Algorithm 1 is bounded in $U \times H$ and every weak cluster point is a solution of (8). Moreover,*

$$\sum_{k \in \mathbb{N}} \|Ku_k + v_k - Y\|^2 + \|K(u_k - u_{k-1})\|^2 < \infty.$$

Remark 2.3. i) Theorem 2.2 implies, that each weak cluster point of $\{u_k\}_{k \geq 1}$ is a solution of (7). In particular, if the solution $u^\dagger$ of (7) is unique (e.g. if J is strictly convex), then $u_k \rightharpoonup u^\dagger$.

ii) Note in particular that (12) is independent of the choice of J, while (13) is independent of the multiresolution statistic being used. This decomposition gives the proposed method a neat modular appeal: once an efficient solution method for the projection problem (12) is established (see e.g. Section 2.3), the regularization functional J in (3) can easily be replaced by providing an algorithm for the penalized least squares problem (13). For most popular choices of J, problem (13) is well studied and efficient computational methods are at hand (see [44] for a extensive collection of algorithms and [32] for an overview on MCMC methods).

For a given tolerance $\tau > 0$, Theorem 2.2 implies that Algorithm 1 terminates and outputs approximate solution $u[\tau]$ and $v[\tau]$ of (8). However, the breaking condition in Algorithm 1

Algorithm 1 Alternating Direction Method of Multipliers

Require: $Y \in H$ (data), $\lambda > 0$ (step size), $\tau \geq 0$ (tolerance).

Ensure: $(u[\tau], v[\tau])$ is an approximate solution of (8) computed in $k[\tau]$ iteration steps.

    $u_0 \leftarrow \mathbf{0}_U$ and $v_0 = p_0 \leftarrow \mathbf{0}_H$

    $r \leftarrow \|Ku_0 + v_0 - Y\|$ and $k \leftarrow 0$.

    while $r > \tau$ **do**

        $k \leftarrow k + 1$.

        $v_k \leftarrow \tilde{v}$ where $\tilde{v} \in \mathcal{C}$ satisfies

$$\|\tilde{v} - (Y + \lambda p_{k-1} - Ku_{k-1})\|^2 \leq \|v - (Y + \lambda p_{k-1} - Ku_{k-1})\|^2 \quad \forall(v \in \mathcal{C}). \qquad (12)$$

        $u_k \leftarrow \tilde{u}$ where $\tilde{u}$ satisfies

$$\frac{1}{2}\|K\tilde{u} - (Y + \lambda p_{k-1} - v_k)\|^2 + \lambda J(\tilde{u}) \leq \frac{1}{2}\|Ku - (Y + \lambda p_{k-1} - v_k)\|^2 + \lambda J(u) \quad \forall(u \in U).$$
$$(13)$$

        $p_k \leftarrow p_{k-1} - (Ku_k + v_k - Y)/\lambda$.

        $r \leftarrow \max(\|Ku_k + v_k - Y\|, \|K(u_k - u_{k-1})\|)$.

    end while

    $u[\tau] \leftarrow u_k$ and $v[\tau] \leftarrow v_k$ and $k[\tau] \leftarrow k$.

---

merely guarantees that the linear constraint in (8) is approximated sufficiently well. Moreover, we know from construction that $v[\tau] \in \mathcal{C}$, which implies $G(v[\tau]) = 0$. So, it remains to evaluate the validity of $u[\tau]$:

Theorem 2.4. *Let $(\hat{u}, \hat{v}, \hat{p}) \in U \times H \times H$ be any saddle point of L_λ. Moreover, let $\tau > 0$ and $u[\tau] \in U$ be returend by Algorithm 1. Then,*

$$0 \leq J(u[\tau]) - J(\hat{u}) - \langle K^*\hat{p}, u[\tau] - \hat{u}\rangle_U \leq \tau\left(6\|\hat{p}\| + \frac{4\|K\hat{u}\| + 2\tau}{\lambda}\right) \quad \forall(\tau > 0).$$

The result in Theorem 2.4 shows how the accuracy of the approximate solution $u[\tau]$ depends on τ. Moreover, it reveals that choosing a small step size λ in Algorithm 1 possibly yields a slow decay of the objective functional J. However, it follows from the definition of L_λ in (10) that a small value for λ fosters the linear constraint in (8).

Corollary 2.5. *Let the assumtions of Theorem 2.4 be satisfied. Moreover, assume that J is a quadratic functional, i.e. $J(u) = \frac{1}{2}\|Lu\|_V^2$, where V is a further Hilbert-space and $L : U \supset D \to V$ is a linear, densely-defined and closed operator. Then*

$$\|L(u[\tau] - \hat{u})\| = \mathcal{O}(\sqrt{\tau}) \quad \forall(\tau > 0).$$

Example 2.6 (Dantzig selector)**.** As already mentioned in the introduction, SMRE (i.e. finding solutions of (3)) reduces to the computation of *Dantzig selectors* for the particular setting $d = 1$, $U = \mathbb{R}^p$ (with usually $p \gg m$) and

$$J(u) = \|u\|_1.$$

When applying Algorithm 1 the subproblem (13) amounts to compute

$$u_k \in \operatorname*{argmin}_{u \in \mathbb{R}^p} \frac{1}{2}\|Ku - (Y + \lambda p_{k-1} - v_k)\|^2 + \lambda\|u\|_1.$$

This is the well known *least absolute shrinkage and selection operator (LASSO)* estimator [43]. For the classical Dantzig selector, one chooses $\mathcal{S} = \{1, \ldots, p\}$ and defines for $S \in \mathcal{S}$ the weight $\omega^S = K\chi_{\{S\}}$. Hence, the subproblem (12) in this case consists in the orthonormal projection of $Y_k = Y + \lambda p_{k-1} - Ku_{k-1}$ onto the set

$$\mathcal{C} = \left\{ v \in \mathbb{R}^m \ : \ \left| \sum_{1 \leq j \leq m} \omega_j^S v_j \right| \leq q \text{ for } 1 \leq S \leq p \right\}.$$

The implications of Theorem 2.4 in the present case are in general rather weak. If the saddle point $\hat{u}$ is known to be S-sparse and when K restricted to the support of $\hat{u}$ is injective, then it can be shown that $|u[\tau] - \hat{u}|_1 = \mathcal{O}(\tau)$.

We finally note that for this particular situation a slightly different decomposition than proposed in (2.2) is favorable. To be more precise, define $\tilde{K} = K^T K$ and $\tilde{Y} = K^T Y$ and consider

$$J(u) + \tilde{G}(v) \quad \text{subject to} \quad \tilde{K}u - v = \tilde{Y}.$$

where $\tilde{G}$ is the characteristic function on the set $\{v \in H \ : \ \|v\|_\infty \leq q\}$. Algorithm 1 applied to this modified decomposition then results in the ADMM as introduced in [33]. In this case the projection in step (12) has a closed from.

2.3. **The Projection Problem.** Algorithm 1 resolves the constrained convex optimization problem (7) into a quadratic program (12) and an unconstrained optimization problem (13). The quadratic program (12) in the k-th step of Algorithm 1 can be written as a projection:

$$\inf_{v \in H} \|v - Y_k\|^2 \quad \text{subject to} \quad \mu_S(v) \leq q \ \forall (s \in \mathcal{S}) \tag{14}$$

where $Y_k = Y + \lambda p_{k-1} - Ku_{k-1}$. We reformulate the side conditions to

$$v \in \mathcal{C} = \bigcap_{S \in \mathcal{S}} C_S \quad \text{where} \quad C_S = \{v \in H : \mu_S(v) \leq q\}. \tag{15}$$

The sets C_S are closed and convex and problem (14) thus amounts to compute the projection $P_{\mathcal{C}}(Y_k)$ of Y_k onto the intersection $\mathcal{C}$ of closed and convex sets. According to this interpretation, we use Dykstra's projection algorithm as introduced in [5] to solve (14). This algorithm takes an element $v \in H$ and convex sets $D_1, \ldots, D_M \subset H$ as arguments. It then creates a sequence converging to the projection of v onto the intersection of the D_j by successively performing projections onto individual D_j's. To this end, let $P_D(\cdot)$ denote the projection onto $D \subset H$ and $S_D = P_D - \text{Id}$ be the corresponding projection step. Dykstra's method is summarized in Algorithm 2.

A natural explanation of the algorithm in a primal-dual framework as well as a proof that the sequence $\{h(M, k)\}_{k \in \mathbb{N}}$ converges to $P_{\mathcal{D}}(h)$ in norm can be found in [13, 23]. For the case when $\mathcal{D}$ constitutes a polyhedron even explicit error estimates are at hand (cf. [46]):

Theorem 2.7. *Let $\{h_k\}_{k \in \mathbb{N}}$ be the sequence generated by Algorithm 2 and $P_{\mathcal{D}}(h)$ be the projection of the input h onto $\mathcal{D}$. Then there exist constants $\rho > 0$ and $0 \leq c < 1$ such that for all $k \in \mathbb{N}$*

$$\|h_k - P_{\mathcal{D}}(h)\| \leq \rho c^k.$$

Remark 2.8. The constant c increases with the number M of convex sets which intersection form the set D that h is to be projected on. The convergence rate therefore improves with decreasing M. For further details and estimates for the constants ρ and c, we refer to [46].

Algorithm 2 Dykstra's Algorithm

Require: $h \in H$ (data), $D_1, \ldots, D_M \subset H$ (closed and convex sets)
Ensure: A sequence $\{h_k\}_{k \in \mathbb{N}}$ that converges strongly to $P_{\mathcal{D}}(h)$ where $\mathcal{D} = \bigcap_{j=1,\ldots,M} D_j$

  $h_{0,0} \leftarrow h$
  for $j = 1$ to M **do**
    $h_{0,j} \leftarrow P_{D_j}(h_{0,j-1})$ and $Q_{0,j} \leftarrow S_{D_j}(h_{0,j-1})$
  end for
  $h_1 \leftarrow h_{0,M}$ and $k \leftarrow 1$
  for $k \geq 1$ **do**
    $h_{k,0} \leftarrow h_k$
    for $j = 1$ to M **do**
      $h_{k,j} \leftarrow P_{D_j}(h_{k,j-1} - Q_{k-1,j})$ and $Q_{k,j} \leftarrow S_{D_j}(h_{k,j-1} - Q_{k-1,j})$
    end for
    $h_{k+1} \leftarrow h_{k,M}$ and $k \leftarrow k + 1$
  end for

Note that application of Dykstra's algorithm is particularly appealing if the projections P_{D_j} can be easily computed or even stated explicitly, as it is the case in the following examples.

Example 2.9. Assume that $\Lambda = \mathrm{Id}$. Then the sets C_S are the rectangular cylinders

$$C_S = \left\{ v \in H \ : \ \left| \langle \omega^S, v \rangle \right| \leq q \right\}.$$

The projection can therefore be explicitly computed as

$$P_{C_S}(v) = \begin{cases} v - \mathrm{sign}\left(\langle \omega^S, v \rangle \right) \frac{\omega^S}{\|\omega^S\|} \left(\frac{|\langle \omega^S, v \rangle|}{\|\omega^S\|} - q \right) & \text{if } \mu_S(v) > q \\ v & \text{else} \end{cases}.$$

The left image in Figure 1 depicts an example for $\mathcal{C}$ for $H = \mathbb{R}^2$. For a detailed geometric interpretation of the MR-statistic we also refer to [36].

Example 2.10. Assume that $\Lambda(v)_{\boldsymbol{\nu}} = v_{\boldsymbol{\nu}}^2$. Then, it follows that $v \mapsto \langle \omega^S, \Lambda(v) \rangle$ is convex if and only if $\omega_{\boldsymbol{\nu}}^S \geq 0$ for all $\boldsymbol{\nu} \in X$. In this case, the sets C_S are elliptic cylinders

$$C_S = \left\{ v \in H \ : \ \sum_{\boldsymbol{\nu} \in X} \omega_{\boldsymbol{\nu}}^S v_{\boldsymbol{\nu}}^2 \leq q \right\}.$$

Moreover, if $\omega_{\boldsymbol{\nu}}^S \in \{0, 1\}$ for all $\boldsymbol{\nu} \in X$, then the projection P_{C_S} can be explicitly computed as

$$P_{C_S}(v) = \begin{cases} \frac{q}{\langle \omega^S, \Lambda(v) \rangle} v \chi_{\{\omega^S = 1\}} + v \chi_{\{\omega^S = 0\}} & \text{if } \mu_S(v) > q \\ v & \text{else} \end{cases}.$$

The right image in Figure 1 depicts an example of $\mathcal{C}$ for $H = \mathbb{R}^2$.

A first approach to use Dykstra's algorithm to solve (14) is to set $M = \#\mathcal{S}$ and identify D_j with C_S for all $j = 1, \ldots, M$. In view of Remark 2.8, however, it is clearly desirable to decrease the number M of convex sets that enter Dykstra's algorithm. In order to do so, we take a slightly more sophisticated approach than the one just presented. We partition the set
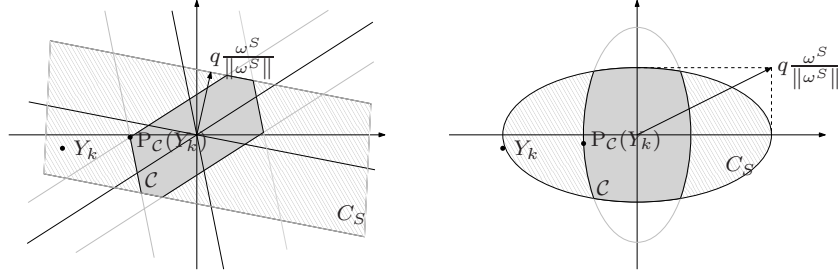
FIGURE 1. Admissible set $\mathcal{C}$ for the projection problem (14) as in Example 2.9 (left) and Example 2.10 (right).

$\mathcal{S}$ into $\mathcal{S}_1, \ldots, \mathcal{S}_M$ such that for all $S \neq \tilde{S} \in \mathcal{S}_j$

$$\omega^S \perp \omega^{\tilde{S}} \quad \text{and} \quad \frac{\partial}{\partial v_{\boldsymbol{\nu}}} \Lambda(\cdot)_{\tilde{\boldsymbol{\nu}}} \equiv 0, \forall (\boldsymbol{\nu} \in S, \tilde{\boldsymbol{\nu}} \in \tilde{S}) \tag{16}$$

and regroup $\{C_S\}_{s \in \mathcal{S}}$ into $\{D_1, \ldots, D_M\}$ with

$$D_j = \bigcap_{S \in \mathcal{S}_j} C_S. \tag{17}$$

Given the projections P_{C_S}, the projection onto D_j can be easily computed: For $v \in H$ identify the set

$$V_j = \{S \in \mathcal{S}_j : \mu_S(v) > q\}$$

of indices for which v violates the side condition (15) and set

$$P_{D_j}(v) = v - \sum_{S \in V_j} \left(P_{\mathcal{C}_S} - \mathrm{Id} \right) v$$

To keep M small, we choose $\mathcal{S}_1 \subset \{1, \ldots, N\}$ as the biggest set such that (16) holds for all $S, \tilde{S} \in \mathcal{S}_1$. We then choose $\mathcal{S}_2 \subset \mathcal{S} \setminus \mathcal{S}_1$ with the same property and continue in this way until all indices are utilized. While this procedure does not necessarily result into M being minimal with the desired property, it still yields a distinct reduction of N in many practical situations. We will illustrate this approach for SMREs in imaging in Section 3.

3. APPLICATIONS

In this section we will illustrate the capability of Algorithm 1 for computing SMREs in some practical situations: in Section 3.1 we will study a simply one-dimensional regression problem as it was also studied in [12], yet with a different penalty function J. In Section 3.2 we illustrate how SMREs performs in image denoising. In both cases we compare our results to other methods. Finally, we will apply the SMRE technique to the problem of image deblurring in confocal fluorescence microscopy in Section 3.3.

Before we study the aforementioned examples, we clarify some common notation. We will henceforth assume that $U = H = (\mathbb{R}^m)^d$ with $d = 1$ (Section 3.1) and $d = 2$ (Sections 3.2 and 3.3), respectively. Moreover, we will employ gradient based regularization functionals of the form

$$J(u) = TV_p(u) := \frac{1}{p} \sum_{\boldsymbol{\nu} \in X} |\mathrm{D}u_{\boldsymbol{\nu}}|_2^p \quad \text{with } p \in \{1, 2\} \tag{18}$$

where $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^d$ and D denotes the forward difference operator defined by

$$(\mathrm{D}u_{\boldsymbol{\nu}})_i = \begin{cases} u_{\boldsymbol{\nu}+\boldsymbol{e_i}} - u_{\boldsymbol{\nu}} & \text{if } 1 \leq \nu_i \leq n-1 \\ 0 & \text{else.} \end{cases}$$

For the case $p = 2$ the minimization problem (13) amounts to solve an implicit time step of the d-dimensional diffusion equation with initial value $(Y + \lambda p_{k-1} - v_k)$ and time step size λ. This can be solved by a simple (sparse) matrix inversion.

For the case $p = 1$, TV_1 is better known as *total-variation semi-norm*. It was shown in [34] (see also [24] for similar results in the continuous setting) that the *taut-string algorithm* (as introduced in [10]) constitutes an efficient solution method for (13) in the case $d = 1$. In the general case $d \geq 1$, we employ the fixed point approach for solving the Euler-Lagrange equations for (13) described in [14] (see also [44, Chap. 8.2.4]). We finally note that the functional TV_1 fails to be differentiable; a fact that leads to serious numerical problems when trying to compute the Euler-Lagrange conditions for (13). Hence, we will use in our simulations a regularized version of TV_1 defined by

$$TV_1^{\beta}(u) = \sum_{\boldsymbol{\nu}\in X} \sqrt{(\mathrm{D}u_{\boldsymbol{\nu}})_i^2 + \beta^2} \tag{19}$$

for a small constant $\beta > 0$.

Evaluation. In order to evaluate the performance of SMREs, we will employ various distance measures between an estimator $\hat{u}$ and the true signal u^0. On the one hand, we will use standard measures such as *mean integrated squared error (MISE)* and the *mean integrated absolute error (MIAE)* which are given by

$$\mathrm{MISE} = \mathbf{E}\left(\frac{1}{m^d}\sum_{\boldsymbol{\nu}\in X}(\hat{u}_{\boldsymbol{\nu}} - u^0_{\boldsymbol{\nu}})^2\right) \quad \text{and} \quad \mathrm{MIAE} = \mathbf{E}\left(\frac{1}{m^d}\sum_{\boldsymbol{\nu}\in X}|\hat{u}_{\boldsymbol{\nu}} - u^0_{\boldsymbol{\nu}}|\right),$$

respectively. On the other hand, we also intend to measure how well an estimator $\hat{u}$ matches the "smoothness" of the true signal u^0, where smoothness is characterized by the regularization functional J. To this end, we introduce the *symmetric Bregman divergence*

$$D_J^{\mathrm{sym}}(\hat{u}, u^0) = \frac{1}{m^d}\sum_{\boldsymbol{\nu}\in X}\left(\nabla J(\hat{u})_{\boldsymbol{\nu}} - \nabla J(u^0)_{\boldsymbol{\nu}}\right)\left(\hat{u}_{\boldsymbol{\nu}} - u^0_{\boldsymbol{\nu}}\right),$$

where ∇J denotes the gradient of the regularization functional J. Clearly, $D_J^{\mathrm{sym}}(\hat{u}, u^0)$ is symmetric and since J is assumed to be convex, also non-negative. However, the symmetric Bregman divergence usually does not satisfy the triangle inequality and hence in general does not define a (semi-) metric on U [9]. The following examples shed some light on how the Bregman divergence incorporates the functional J in order to measures the distance of $\hat{u}$ and u^0.

Example 3.1. Let $J(u) = TV_p$ as in (18).

i) If $p = 2$, then

$$D_J^{\mathrm{sym}}(\hat{u}, u^0) = \sum_{\boldsymbol{\nu}\in X}\left|\mathrm{D}\hat{u}_{\boldsymbol{\nu}} - \mathrm{D}u^0_{\boldsymbol{\nu}}\right|_2^2.$$

In other words, the symmetric Bregman distance w.r.t. to TV_2 is the mean squared distance of the *derivatives* of $\hat{u}$ and u^0.

ii) If $p = 1$, then

$$
\begin{aligned}
D_J^{\text{sym}}(\hat{u}, u^0) &= \frac{1}{m^2} \sum_{\boldsymbol{\nu} \in X} \left(\frac{\mathrm{D}\hat{u}_{\boldsymbol{\nu}}}{|\mathrm{D}\hat{u}_{\boldsymbol{\nu}}|} - \frac{\mathrm{D}u_{\boldsymbol{\nu}}^0}{|\mathrm{D}u_{\boldsymbol{\nu}}^0|} \right) \cdot \left(\mathrm{D}\hat{u}_{\boldsymbol{\nu}} - \mathrm{D}u_{\boldsymbol{\nu}}^0 \right) \\
&= \frac{1}{m^2} \sum_{\boldsymbol{\nu} \in X} \left(|\mathrm{D}\hat{u}_{\boldsymbol{\nu}}| + |\mathrm{D}u^0| \right) \left(1 - \frac{\mathrm{D}\hat{u}_{\boldsymbol{\nu}}}{|\mathrm{D}\hat{u}_{\boldsymbol{\nu}}|} \cdot \frac{\mathrm{D}u_{\boldsymbol{\nu}}^0}{|\mathrm{D}u_{\boldsymbol{\nu}}^0|} \right) \\
&= \frac{1}{m^2} \sum_{\boldsymbol{\nu} \in X} \left(|\mathrm{D}\hat{u}_{\boldsymbol{\nu}}| + |\mathrm{D}u^0| \right) \left(1 - \cos \gamma_{\boldsymbol{\nu}} \right),
\end{aligned}
$$

where $\gamma_{\boldsymbol{\nu}}$ denotes the angle between the level lines of $\hat{u}$ and u^0 at the point $x_{\boldsymbol{\nu}}$. Put differently, the symmetric Bregman divergence w.r.t the total variation semi-norm TV_1 is small if for sufficiently many points $x_{\boldsymbol{\nu}}$ either both $\hat{u}$ and u^0 are constant in a neighborhood of $x_{\boldsymbol{\nu}}$ or the level lines of $\hat{u}$ and u^0 through $x_{\boldsymbol{\nu}}$ are parallel. In practice rather TV_1^{β} in (19) (for a small $\beta > 0$) instead of TV_1 is used in order to avoid singularities. Then, the above formulas are slightly more complicated.

We will use the mean symmetric Bregman divergence (MSB) given by

$$
\mathrm{MSB} = \mathbf{E} \left(D_J^{\text{sym}}(\hat{u}, u^0) \right)
$$

as an additional evaluation method. In all our simulations we approximate the expectations above by the empirical means of 500 trials.

Comparison with other methods. We will compare the SMREs to other regression methods. Firstly, we will consider estimators obtained by the *global* penalized least squares method:

$$
\hat{u}(\lambda) := \operatorname*{argmin}_{u \in H} \frac{1}{2} \sum_{\boldsymbol{\nu} \in X} (u_{\boldsymbol{\nu}} - Y)^2 + \lambda J(u), \quad \lambda > 0. \tag{20}
$$

In particular, we focus on estimators $\hat{u}(\lambda)$ that are closest (in some sense) to the true function u^0. We call such estimators *oracles*. We define the L^2- and Bregman-oracle by $\hat{u}_{\mathrm{L}^2} = \hat{u}(\lambda_2)$ and $\hat{u}_{\mathrm{B}} = \hat{u}(\lambda_{\mathrm{B}})$, where

$$
\lambda_2 := \mathbf{E} \left(\operatorname*{argmin}_{\lambda > 0} \left\| u^0 - \hat{u}(\lambda) \right\| \right) \quad \text{and} \quad \lambda_{\mathrm{B}} := \mathbf{E} \left(\operatorname*{argmin}_{\lambda > 0} D_J^{\text{sym}}(u^0, \hat{u}(\lambda)) \right)
$$

respectively. Of course, oracles are not available in practice, since the true signal u^0 is unknown. However, they represent ideal instances within the class of estimators given by (20) that usually perform better than any data-driven parameter choices (such as cross-validation) and hence may serve as a reference.

Secondly, we also compare our approach to *adaptive weights smoothing (AWS)* [40] which constitutes a benchmark technique for data-driven, spatially adaptive regression. We compute these estimators by means of the official R-package[2] and denote them by $\hat{u}_{\text{aws}}^{\text{ker}}$, where ker $\in$ {Gaussian, Triangle} decodes the shape of the underlying regression kernel.

---

[2]available at`http://cran.r-project.org/web/packages/aws/index.html`

3.1. **Non-parametric Regression.** In this section we apply the SMRE technique to a non-parametric regression problem in $d = 1$ dimensions, i.e. the noise model (1) becomes

$$Y_{\boldsymbol{\nu}} = u^0_{\boldsymbol{\nu}} + \varepsilon_{\boldsymbol{\nu}} \quad \boldsymbol{\nu} = 1, \ldots, m, \tag{21}$$

where we assume that $\varepsilon_{\boldsymbol{\nu}}$ are independently and normally distributed r.v. with $\mathbf{E}\left(\varepsilon_{\nu}\right) = 0$ and $\mathbf{E}\left(\varepsilon^2_{\nu}\right) = \sigma^2$. The upper left image in Figure 2 depicts the true signal u^0 (solid line) and the data Y, with $m = 1024$ and $\sigma = 0.5$. The application we have in mind with this example arises in NMR spectroscopy, where the NMR spectra provide structural information on the number and type of chemical entities in a molecule. In this context, we suggest to choose $J = TV_2$, since the true signal u^0 is rather smooth (see [12] for examples where J is chosen to be the total variation of the first and second derivative).

Finally, we discuss the MR-statistic T in (4). We choose $\Lambda = \mathrm{Id}$ and the index set $\mathcal{S}$ to consist of all discrete intervals with side lengths ranging from 1 to 100. For an interval $S \in \mathcal{S}$ we set $\omega^S = (\#S)^{-1/2}\chi_S$. Thus, each SMRE solves the constrained optimization problem

$$\inf_{u \in U} TV_2(u) \quad \text{s.t.} \quad \frac{1}{\sqrt{\#S}}\left|\sum_{\boldsymbol{\nu} \in S}(Y - u)_{\boldsymbol{\nu}}\right| \leq q \quad \forall(S \in \mathcal{S}). \tag{22}$$

We choose q to be the α-quantile of the MR-statistic T, that is

$$q_\alpha = \inf\left\{q \in \mathbb{R} \; : \; \mathbb{P}\left(\max_{S \in \mathcal{S}}\frac{1}{\sqrt{\#S}}\left|\sum_{\boldsymbol{\nu} \in S}\varepsilon_{\boldsymbol{\nu}}\right| \leq q\right) \geq \alpha\right\} \quad \alpha \in (0, 1). \tag{23}$$

We note that except for few special cases (cf. [28, 30]) closed form expressions for the distribution of the MT-statistic T are usually not at hand. In practice one rather considers the empirical distribution of T where the variance σ^2 can be estimated at a rate $\sqrt{md}$ (cf. [38]).

We will henceforth denote by $\hat{u}_\alpha$ a solution of (22) with $q = q_\alpha$. As argued in Section 1.1, $\hat{u}_\alpha$ is smoother (i.e. has smaller value TV_2) than the true signal u^0 with a probability of at least α while it satisfies the constraint that the multiresolution statistic T does not exceed q_α. This is a sound statistical interpretation of the regularization parameter α.

Numerical results and simulations. In Figure 2 the oracles $\hat{u}_{\mathrm{L}^2}$ and $\hat{u}_{\mathrm{B}}$, the AWS-estimators $\hat{u}^{\mathrm{Triangle}}_{\mathrm{aws}}$ and and $\hat{u}^{\mathrm{Gauss}}_{\mathrm{aws}}$ as well as the SMRE $\hat{u}_{0.9}$, $\hat{u}_{0.75}$ and $\hat{u}_{0.5}$ are depicted. It is evident that the SMRE matches the smoothness of the true object much better than the other estimators while the essential features of the signal (such as peak location and peak height) are preserved. In particular, almost no additional local extrema are generated by our approach which stays in obvious contrast to the other methods. Moreover, we point out that the SMRE are quite robust w.r.t. the choice of the confidence level α.

We verify this behavior by a simulation study in Table 1. For different noise levels ($\sigma = 0.1, 0.3$ and $0, 5$) we compare the MISE, MIAE and MSB. Additionally, we compute the *mean number of local maxima (MLM)* of $\hat{u}$ relative to the number of local maxima in u^0 (which is 11). Here $\hat{u}$ is any of the above estimators. Note that the latter measure (similar to the MSB) takes into account the smoothness of the estimators where a value MLM $\gg 1$ indicates too many local maxima and hence a lack of regularity whereas MLM < 1 implies severe oversmoothing.

As it becomes apparent from Table 1, the SMREs are performing similarly well when compared to the reference estimators as far as the standard measures MISE and MIAE are concerned. For small noise levels ($\sigma = 0.1$) SMREs even prove to be superior. The distance measures MSB and MLM, however, are significantly smaller for SMREs which indicates that
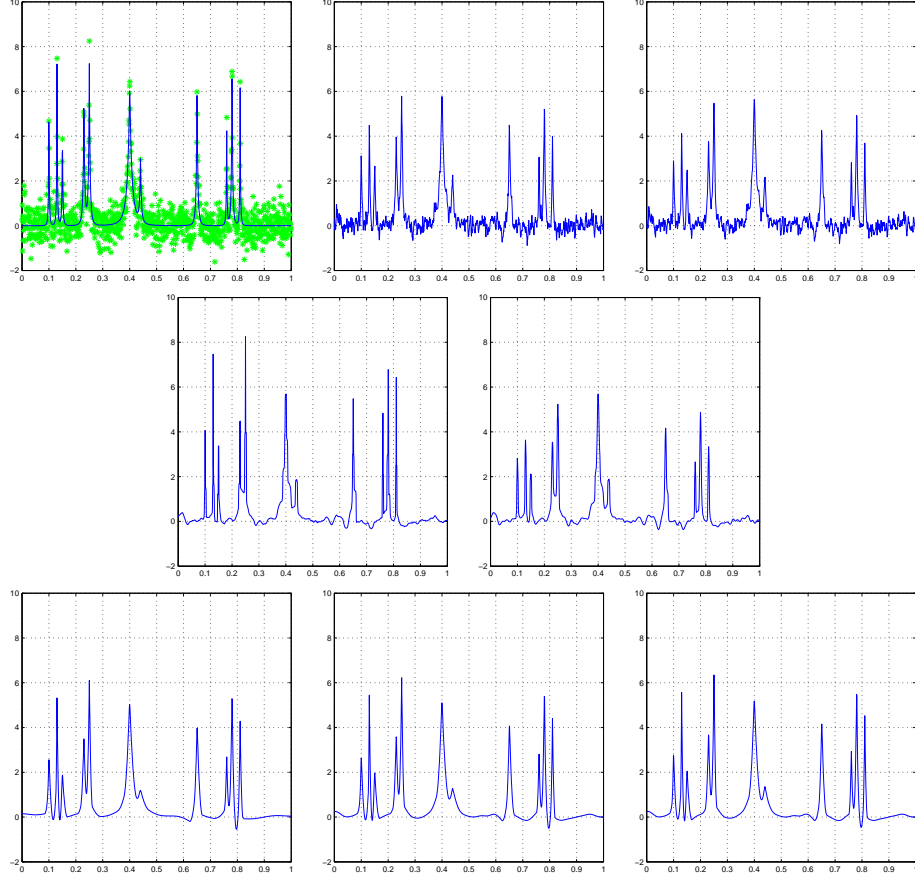
FIGURE 2. Row-wise from top left: true signal u^0 (solid) with data Y, oracles $\hat{u}_{\mathrm{L}^2}$ and $\hat{u}_{\mathrm{B}}$, AWS estimators $\hat{u}_{\mathrm{aws}}^{\mathrm{Triangle}}$ and $\hat{u}_{\mathrm{aws}}^{\mathrm{Gaussian}}$, and the SMREs $\hat{u}_{0.9}$, $\hat{u}_{0.75}$ and $\hat{u}_{0.5}$.

these meet the smoothness of the true object u^0 much better than the reference estimators (cf. Example 3.1 i)). All in all, the simulation results confirm our visual impressions above.

Implementation Details. The current index set $\mathcal{S}$ results in an overall number of constraints in (22) of

$$\#\mathcal{S} = \sum_{i=1}^{100}(1024 - i + 1) = 97450.$$

As pointed out in Section 2.3, the efficiency of Dykstra's Algorithm can be increased by grouping independent side-conditions, that is side-conditions corresponding to intervals in $\mathcal{S}$ with empty intersection. For example, the system $\mathcal{S}$ can be grouped such that the intersection of the corresponding sets $D_1, \ldots, D_M$ in (17) form $\mathcal{C}$ with

$$M = \sum_{i=1}^{100} i = 5050.$$

	$\sigma = 0.1$				$\sigma = 0.3$			
	MISE	MSB	MIAE	MLM	MISE	MSB	MIAE	MLM
$\hat{u}_{\mathrm{L}^2}$	0.009	0.008	0.071	11.881	0.046	0.027	0.156	10.915
$\hat{u}_{\mathrm{B}}$	0.009	0.007	0.070	11.700	0.048	0.026	0.149	10.359
$\hat{u}_{\mathrm{aws}}^{\mathrm{Triangle}}$	0.007	0.007	0.048	2.551	0.040	0.035	0.112	3.053
$\hat{u}_{\mathrm{aws}}^{\mathrm{Gauss}}$	0.054	0.040	0.068	1.971	0.062	0.041	0.107	2.230
$\hat{u}_{0.9}$	0.008	0.004	0.047	1.336	0.056	0.019	0.127	1.273
$\hat{u}_{0.75}$	0.007	0.004	0.044	1.342	0.050	0.018	0.121	1.290
$\hat{u}_{0.5}$	0.007	0.004	0.043	1.366	0.046	0.017	0.116	1.290

	$\sigma = 0.5$			
	MISE	MSB	MIAE	MLM
$\hat{u}_{\mathrm{L}^2}$	0.091	0.037	0.213	9.860
$\hat{u}_{\mathrm{B}}$	0.094	0.036	0.206	9.135
$\hat{u}_{\mathrm{aws}}^{\mathrm{Triangle}}$	0.078	0.058	0.162	3.141
$\hat{u}_{\mathrm{aws}}^{\mathrm{Gauss}}$	0.079	0.043	0.149	2.330
$\hat{u}_{0.9}$	0.134	0.034	0.207	1.194
$\hat{u}_{0.75}$	0.120	0.032	0.196	1.241
$\hat{u}_{0.5}$	0.109	0.030	0.186	1.238

TABLE 1. Simulation studies for one dimensional peak data set.

In all our simulations we set $\tau = 10^{-4}$ and $\lambda = 1.0$ in Algorithm 1 which results in $k[\tau] \approx 100$ iterations and an overall computation time of approximately 20 minutes for each SMRE. We note, however, that more than 95% of the computation time is needed for the projection step (12) and that a considerable speed up for the latter could be achieved by parallelization.

3.2. **Image denoising.** In this section we apply the SMRE technique to the problem of image denoising, that is non-parametric regression in $d = 2$ dimensions. In other words, we consider the noise model (21) as in Section 3.1, where the index $\boldsymbol{\nu}$ ranges over the discrete square $\{1, \ldots, m\}^2$. In Figure 3 two typical examples for images u^0 and noisy observations Y are depicted ($m = 512$ and $\sigma = 0.1$, where u^0 is scaled between 0 (black) and 1 (white)).

We will use the total-variation semi-norm $J = TV_1^{\beta}$ as regularization functional ($\beta = 10^{-8}$). Moreover, we choose Λ to be defined as

$$\Lambda(v)_{\boldsymbol{\nu}} = v_{\boldsymbol{\nu}}^2, \quad \forall (\boldsymbol{\nu} \in 1, \ldots, m^2). \tag{24}$$

The index set $\mathcal{S}$ is defined to be the collection of all discrete squares with side lengths ranging from 1 to 25 and we set $\omega^S = c_S \chi_S$ with yet to be defined constants c_S. Thus, each SMREs solves the constrained optimization problem

$$\inf_{u \in U} TV_1^{\beta}(u) \quad \text{s.t.} \quad \sum_{\boldsymbol{\nu} \in S} c_S (Y - u)_{\boldsymbol{\nu}}^2 \leq q \quad \forall (S \in \mathcal{S}). \tag{25}$$

We agree upon $q = 1$ and specify the constants c_S. To this end, compute for $s = 1, \ldots, 25$ the quantile values

$$q_{\alpha,s} = \inf \left\{ q \in \mathbb{R} \ : \ \mathbb{P} \left(\max_{\substack{S \in \mathcal{S} \\ \#S = s}} \sum_{\boldsymbol{\nu} \in S} \varepsilon_{\boldsymbol{\nu}}^2 \leq q \right) \geq 1 - \alpha \right\} \quad \alpha \in (0, 1)$$

and set $c_S = q_{\alpha, \#S}^{-1}$. In other words, the definition of c_S implies that the true signal u^0 satisfies the constraints in (25) *for squares of a fixed side length s* with probability at least α. We will henceforth denote by $\hat{u}_\alpha$ a solution of (25). We remark on this particular choice of the parameters ω_S below.

FIGURE 3. Standard test images "cameraman" (top) and "roof" (bottom) and their noisy counterparts.

Numerical results and simulations. In Figures 4 and 5 the oracles $\hat{u}_{L^2}$ and $\hat{u}_B$, the AWS-estimators $\hat{u}_{aws}^{Triangle}$ and $\hat{u}_{aws}^{Gauss}$ as well as the SMRE $\hat{u}_{0.9}$ are depicted (for the "cameraman" and "roof" test image respectively). It is rather obvious that the L^2-oracles are not favorable: although relevant details in the image are preserved, smooth parts (as e.g. the sky) still contain random structures. In contrast, the estimator $\hat{u}_{aws}^{Gauss}$ preserves smooth areas but looses essential details. The aws-estimator with triangular kernel performs much better, however, it gives piecewise constant reconstructions of smoothly varying portions of the image, which is clearly undesirable. The SMRE and the Bregman-oracle visually perform superior to the other methods. The good performance of the Bregman-oracle indicates that the symmetric Bregman distance is a good measure for comparing images. In contrast to the Bregman-oracle, the SMRE adapts the amount of smoothing to the underlying image structure: constant image areas are smoothed nicely (e.g. sky portions), while oscillating patterns (e.g. the grass part in the "cameraman" image or the roof tiles in the "roof" image) are recovered.

We evaluate the performance of the SMREs by means of a simulation study. To this end, we compute the MISE, MIAE and MSB and compare these values with the reference estimators. We note, however, that in particular the MISE and MIAE are not well suited in order to measure the distance of images for they are inconsistent with human eye perception. In [45] the *structural similarity index (SSIM)* was introduced for image quality assessment that takes into account luminance, contrast and structure of the images at the same time. We use the
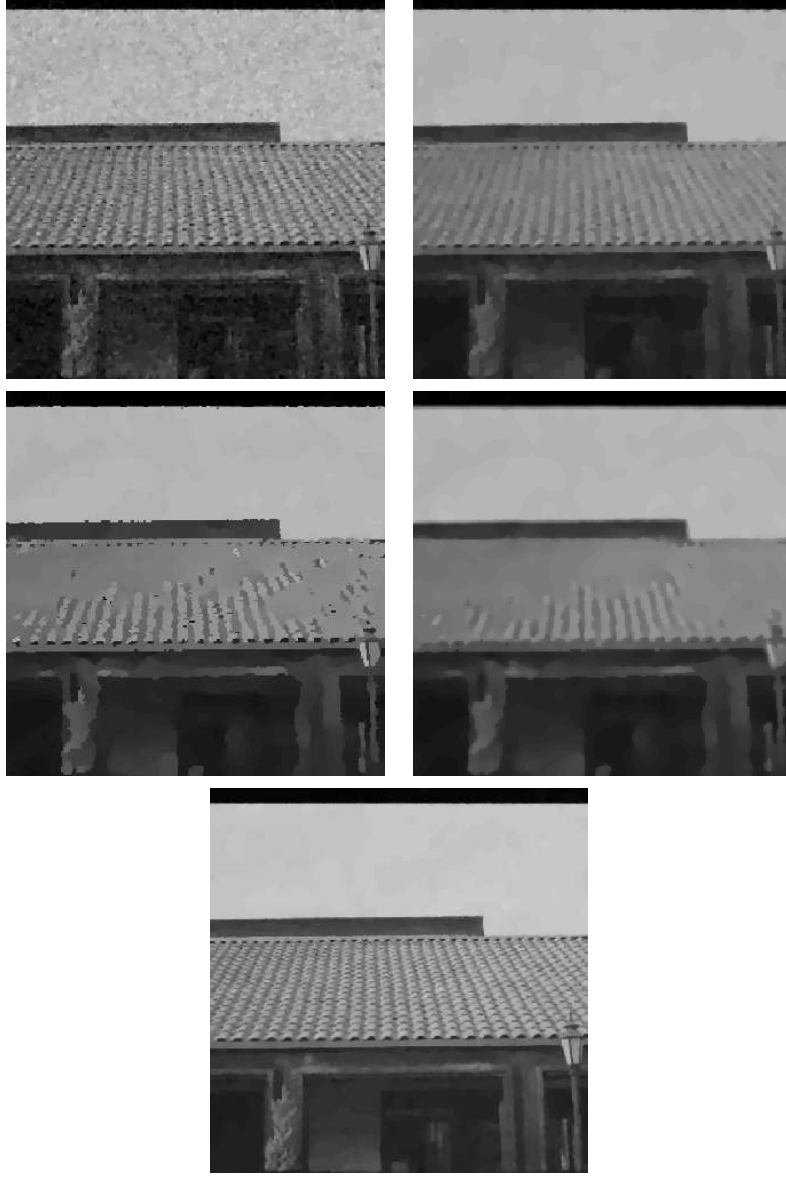
FIGURE 4. Reconstructions "cameraman" (row-wise from top left): L^2-oracle $\hat{u}_{L^2}$, Bregman-oracle $\hat{u}_B$, AWS estimators $\hat{u}_{\text{aws}}^{\text{Triangle}}$ and $\hat{u}_{\text{aws}}^{\text{Gaussian}}$, and SMRE $\hat{u}_{0.9}$.

author's implementation [3] which is normalized such that the SSIM lies in the interval $[-1, 1]$ and is 1 in case of a perfect match. We denote by MSSIM the empirical mean of the SSIM in our simulations.

In Table 2 the simulation results are listed. A first striking fact is the good performance of the L^2-oracle w.r.t. the MISE and MIAE which is supposed to imply reconstruction properties

---

[3]available at https://www.ece.uwaterloo.ca/~z70wang/research/ssim/

FIGURE 5. Reconstructions "roof" (row-wise from top left): L^2-oracle $\hat{u}_{L^2}$, Bregman-oracle $\hat{u}_B$, AWS estimators $\hat{u}_{\text{aws}}^{\text{Triangle}}$ and $\hat{u}_{\text{aws}}^{\text{Gaussian}}$, and SMRE $\hat{u}_{0.9}$.

superior to the other methods. Keeping in mind the visual comparison in Figures 4 and 5, however, this is rather questionable. On the other hand, it becomes evident that the L^2-oracle has a rather poor performance w.r.t. the MSB which is more suited for measuring image distances. It is therefore remarkable that the SMRE performs equally good as the Bregman-oracle which, in contrast to the SMRE, is not accessible (since u^0 is usually unknown). As far as the structural similarity measure MSSIM is concerned our approach proves to be superior to all others. Finally, the simulation results indicate that aws estimation is not favourable for denoising of natural images.

	"cameraman"				"roof"			
	MISE	MSB	MIAE	MSSIM	MISE	MSB	MIAE	MSSIM
$\hat{u}_{\mathrm{L}^2}$	0.0017	0.0314	0.0276	0.7739	0.0029	0.0499	0.0383	0.6700
$\hat{u}_{\mathrm{B}}$	0.0023	0.0256	0.0275	0.7995	0.0038	0.0405	0.0391	0.6607
$\hat{u}_{\mathrm{aws}}^{\mathrm{Triangle}}$	0.0032	0.0482	0.0308	0.7657	0.0046	0.0702	0.0416	0.6205
$\hat{u}_{\mathrm{aws}}^{\mathrm{Gauss}}$	0.0046	0.0470	0.0360	0.7284	0.0053	0.0686	0.0457	0.5668
$\hat{u}_{0.9}$	0.0021	0.0252	0.0297	0.8024	0.0033	0.0374	0.0407	0.7003

TABLE 2. Simulation studies for the test images "cameraman" and "roof".

Notes on the choice of Λ and ω^S. In general, a proper choice of the transformation Λ and of the weight-functions ω^S can be achieved by including prior structural information on the true image to be estimated. Substantial parts of natural images, such as photographs, consists of oscillating patterns (as e.g. fabric, wood, hair, grass etc.). This becomes obvious in the standard test images depicted in Figure 3. We claim that for signals that exhibit oscillating patterns, a quadratic transformation Λ as in (24) is favorable, since it yields (compared to the linear statistic studied in Section 3.1) a larger power of the local test statistic on small scales.

In order to illustrate this, we simulate noisy observations Y of the test images u in Figure 3 as in (21) with $\sigma = 0.1$ and compute a *global* estimator $\hat{u}$ by computing a minimizer of the ROF-functional (20) (with $\lambda = 0.1$). We intend to examine how well over-smoothed regions in $\hat{u}$ are detected by the MR-statistic $T(Y - \hat{u})$ as in (4) with two different average functions (cf. (5))

$$\mu_{1,S}(v) = \left| \sum_{\boldsymbol{\nu} \in S} v_\nu \right| \quad \text{and} \quad \mu_{2,S}(v) = \sum_{\boldsymbol{\nu} \in S} v_\nu^2$$

respectively. For the sake of simplicity we restrict for the moment our considerations on the index set $\mathcal{S}$ of all 5×5 sub-squares in $\{1, \ldots, m\}^2$. In Figure 6 the local means $\mu_{i,S}$ of the residuals $v = Y - \hat{u}$ for the "roof"-image are depicted. To be more precise, the center coordinate of each square $S \in \mathcal{S}$ is colored according to $\mu_{i,S}$, Hence, large values indicate locations where the estimator $\hat{u}$ is considered over-smoothed according to the statistic. It becomes visually clear that the localization of oversmoothed regions is better for $\mu_{2,S}$. This is a good motivation for incorporating the local means of the squared residuals in the SMRE model (7).
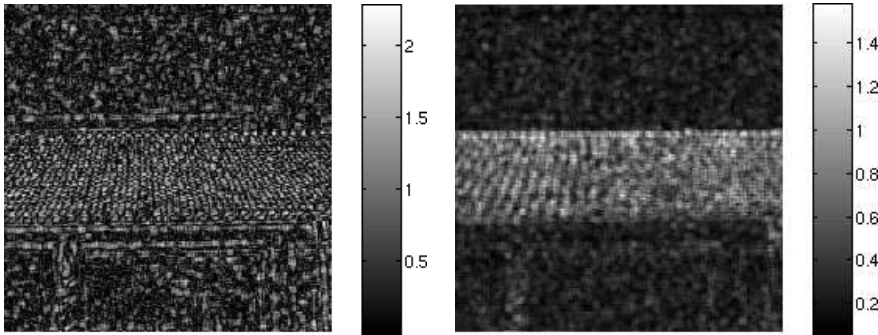
FIGURE 6. Local means $\mu_{1,S}$ (left) and $\mu_{2,S}$ (right)) of the residuals for "roof" image.

We finally comment on the choice of c_S. Since $\varepsilon_{\boldsymbol{\nu}}$ are independent and normally distributed random variables, the (scaled) average function

$$\sigma^{-2}\mu_S(\varepsilon) = \sum_{\boldsymbol{\nu}\in S}\left(\frac{\varepsilon_{\boldsymbol{\nu}}}{\sigma}\right)^2$$

is χ^2 distributed with $\#S$ degrees of freedom. Note that the distribution of $\sigma^{-2}\mu_S(\varepsilon)$ is identical only for sets S of the same scale $\#S$. As a consequence of this, it is likely that certain scales dominate the supremum in the MR-statistic T which spoils the multiscale properties of our approach. As a way out, we compute normalizing constants *for each scale separately.*

An alternative approach would be to search for transformations that turn $\mu_S(\varepsilon)$ into almost identically distributed random variables. Logarithmic and p-root transformations are often employed for this purpose (see e.g. [25]). This will be investigated separately.

Implementation Details. The current index set $\mathcal{S}$ results in an overall number of constraints in (25) of

$$\#\mathcal{S} = \sum_{i=1}^{25}(512 - i + 1)^2 = 6251300.$$

Again by grouping independent side-conditions, the system $\mathcal{S}$ can be grouped such that the intersection of the corresponding sets $D_1, \ldots, D_M$ in (17) form $\mathcal{C}$ with

$$M = \sum_{i=1}^{25} i^2 = 5525.$$

In all our simulations we set $\tau = 10^{-4}$ and $\lambda = 0.25$ in Algorithm 1 which results in $k[\tau] \approx 30$ iterations and a overall computation time of approximately 2 hours for each SMRE. Hence, parallelization is clearly desirable in this case.

3.3. **Deconvolution.** Another interesting class of problems which can be approached by means of SMREs are deconvolution problems. To be more precise, we assume that K is a convolution operator, that is

$$(Ku)_{\boldsymbol{\nu}} = (k * u)_{\boldsymbol{\nu}} = \sum_{\boldsymbol{m}\in\mathbb{R}^d} k_{\boldsymbol{\nu}-\boldsymbol{m}}u_{\boldsymbol{m}}$$

where k is a square-summable kernel on the lattice $\mathbb{Z}^d$ and $u \in H$ is extended by zero-padding. We will focus on the situation where k is a circular Gaussian kernel with standard deviation σ given by

$$k_{\boldsymbol{\nu}} = \frac{1}{(\sqrt{2\pi}\sigma)^d}e^{-\frac{\sum_{i=1}^{d}\nu_i^2}{2\sigma^2}}. \tag{26}$$

With $Z = Y + \lambda p_{k-1} + v_k$, the primal step (13) in Algorithm 1 amounts to solve

$$u_k \leftarrow \underset{u\in H}{\operatorname{argmin}} \frac{1}{2}\sum_{\boldsymbol{\nu}\in X}((Ku)_{\boldsymbol{\nu}} - Z_{\boldsymbol{\nu}})^2 + \lambda J(u),$$

where we choose J to be as in (18) and apply the techniques described in [44] for the numerical solution.

In order to illustrate the performance of our approach in practical applications, we give an example from confocal microscopy, nowadays a standard technique in fluorescence microscopy

(cf. [39]). When recording images with this kind of microscope, the original object gets blurred by a Gaussian kernel (in first order). The observations (photon counts) can be modeled as a Poisson process, i.e.

$$Y_{\boldsymbol{\nu}} = \mathrm{Poiss}((Ku^0)_{\boldsymbol{\nu}}), \quad \boldsymbol{\nu} \in X. \tag{27}$$

The image depicted in Figure 7(a) shows a recording of a PtK2 cell taken from the kidney of *potorous tridactylus*. Before the recording, the protein β-tubulin was tagged with a fluorescent marker such that it can be traced by the microscope. The image in 7(a) shows an area of 18×18 µm^2 at a resolution of 798×798 pixel. The point spread function of the optical system can be modeled as a Gaussian kernel with full width at half maximum of 230nm, which corresponds to $\sigma = 4.3422$ in (26).

Note that (27) does not fall immediately into the range of models covered by (1). We will adapt the present situation to the SMRE methodology described in Section 1 by standardization and consider instead of (3) the modified problem

$$\inf_{u \in U} J(u) \quad \text{s.t.} \quad T\left(\frac{Y - Ku}{\sqrt{Ku}}\right) \leq 1 \tag{28}$$

where the division is understood pointwise. Clearly, the problem of finding a solution of (28) is much more involved than solving (3) for the constraints being *nonconvex*: firstly, the functional G as defined in (9) is nonconvex as a consequence of which the convergence result in Theorem 2.2 does not apply and secondly Dykstra's projection algorithm as described in Section 2 cannot be employed.

We propose the following ansatz in order to circumvent this problem: instead of projecting onto the intersection $\mathcal{C}$ of sets C_S as described in (15), we now project in the k-th step of Algorithm 1 onto

$$\mathcal{C}_P[k] = \bigcap_{n=1}^{N} C_{P,S}[k] \quad \text{where} \quad C_{P,S}[k] = \left\{v \in H \ : \ \mu_S\left(v/\sqrt{Ku_k}\right) \leq q\right\}.$$

with a pointwise division by the square root of Ku_k. Put differently, in the k-th step of Algorithm 1 we use the previous estimate u_k of u^0 as a *lagged standardization* in order to approximate the constraints in (28). In fact, we use $\sqrt{\max(Ku_k, \varepsilon)}$ with a small number $\varepsilon > 0$ for standardization, in order to avoid instabilities.

We note that while with this modification Dykstra's algorithm becomes applicable again, the projection problem (12) now changes in each iteration step of Algorithm 1. As a consequence, Theorem 2.2 does not hold anymore after this modification, either. So far, we have not come up with a similar convergence analysis.

We compute the SMRE $\hat{u}_{0.9}$ by employing Algorithm 1 with the modifications described above. As in the denoising examples in Section 3.2 the index set $\mathcal{S}$ consists of all squares with the side-lengths $\{1, \ldots, 25\}$ and we choose $\omega^S = \chi_S$ and $\Lambda = \mathrm{Id}$. We note, that this results in an overall number of $\#\mathcal{S} = 95\ 436\ 200$ inequality constraints. The constant q are chosen as in (23), where we assume that $\varepsilon_{\boldsymbol{\nu}}$ are independent and standard normally distributed r.v.

In Algorithm 1 we set $\lambda = 0.05$ and compute 100 steps. We observe that after a few iterations (~ 15) the error τ falls below 10^{-3} and almost stagnates thereafter, which is due to the fact that we do not increase the accuracy in the subroutines for (13) and (12). Each iteration step in Algorithm 1 approximately takes 10 minutes, where 90% of the computation time is needed for (13). The result is depicted in Figure 7(b).
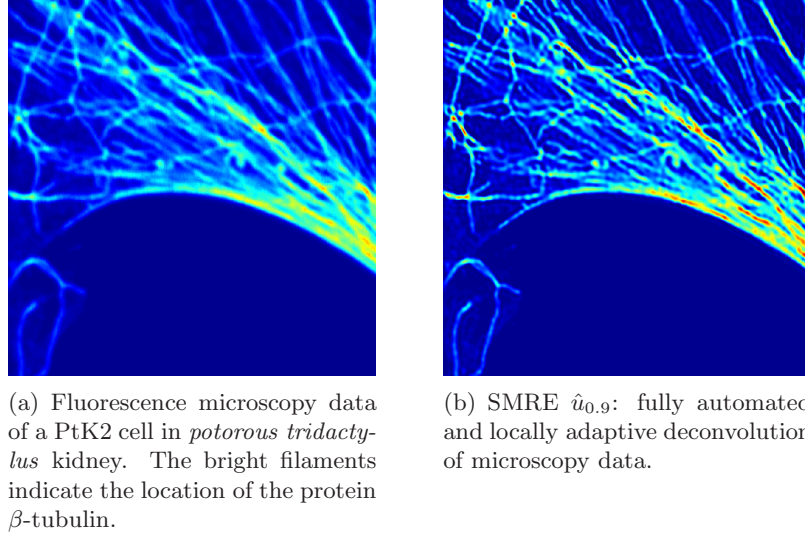
The benefits of our method are twofold:

(a) Fluorescence microscopy data of a PtK2 cell in *potorous tridactylus* kidney. The bright filaments indicate the location of the protein β-tubulin.

(b) SMRE $\hat{u}_{0.9}$: fully automated and locally adaptive deconvolution of microscopy data.

FIGURE 7. Reconstruction of confocal microscopy data.

i) The amount of regularization is chosen in a *completely automatic way*. The only parameter to be selected is the level α in (23). Note that the parameter λ in Algorithm 1 has no effect on the output (though it has an effect on the number of iterations needed and the numerical stability).

ii) The reconstruction has an appealing locally adaptive behavior which in the present example mainly concerns the gaps between the protein filaments: whereas the marked β-tubulin is concentrated in regions of basically one scale, the gaps in between actually make up the multiscale nature of the image.

(a) STED microscopy recording of the PtK2 cell data set.

(b) Detail comparison between confocal recording (left), SMRE $\hat{u}_{0.9}$ (middle) and STED recording (right).
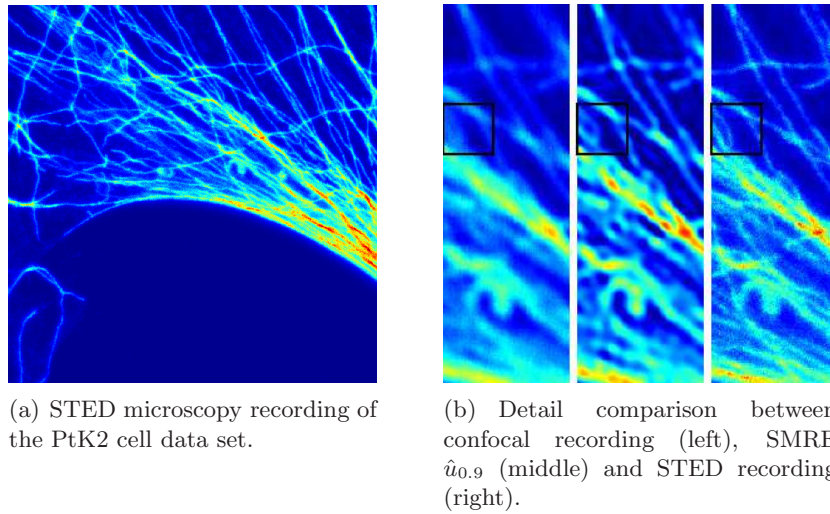
FIGURE 8. Comparison with high-resolution STED microscopy data.

In the present situation we are in the comfortable position to have a reference image at hand by means of which we can evaluate the result of our method: STED (STimulated Emission Depletion) microscopy constitutes a relatively new method, that is capable of recording images at a physically 5-10 times higher resolution as confocal microscopy (see [26, 27]). Hence a STED image of this object may serve as "gold standard" reference image.

Figure 8(a) depicts a STED recording of the PtK2 cell data set in Figure 7(a). The comparison of the SMRE $\hat{u}_{0.9}$ with the STED recording in Figure 8(b) shows that our SMRE technique chooses a reasonable amount of regularization: no artifacts due to under-regularization are generated and on the other hand almost all relevant geometrical features that are present in the high-resolution STED recording become visible in the reconstruction. In particular, we note that filament bifurcations (one such bifurcation is marked by a black box in Figure 8(b)) become apparent in our reconstruction that are not visible in the recorded data.

Finally, we mention that aside to standardization, other transformations of the Poisson data (27) could possibly be considered. For example *Anscombe's transformation* is known to yield reasonable approximations to normality even for low Poisson-intensities and hence has a particular appeal for e.g. microscopy data with low photon-counts. We are currently investigating SMREs that employ Anscombe's transform, where in particular the arising projection problems are challenging.

4. Conclusion and Outlook

In this work, we propose a general estimation technique for nonparametric inverse regression problems in the white noise model (1) based on the convex program (7). It amounts to finding a minimizer of a convex regularization functional $J(u)$ over a set of feasible estimators that satisfy the fidelty condition $T(Y - Ku) \leq q$, where T is assumed to be the maximum over simple convex constraints and q is some quantile of the statistic $T(\varepsilon)$. Any such minimizer we call *statistical multiresolution estimator (SMRE)*. This approach covers well known uni-scale techniques, such as the Dantzig selector, but with a vast field of potentially new application areas, such as locally adaptive imaging. The particular appeal of the multi-scale generalization arises for those situations where a "neighboring relationship" within the signal can be employed to gain additional information by "averaging" neighboring residuals. We demonstrate in various examples that this improvement is drastic.

We approach the numerical solution of the problem by the ADMM (cf. Algorithm 1) that decomposes the problem into two subproblems: A J-penalized least squares problem, independent of T, and an orthogonal projection problem onto the feasible set of (7) that is independent of J. The first problem is well studied and for most typical choices of J fast and reliable numerical approaches are at hand. The projection problem, however, is computational demanding, in particular for image denoising applications. We propose Dykstra's cyclic projection method for its approximate solution. Finally, by extensive numerical experiments, we illustrate the performance of our estimation scheme (in nonparametric regression, image denoising and deblurring problems) and the applicability of our algorithmic approach.

Summarizing, this paper is meant to introduce a novel class of statistical estimators, to provide a general algorithmic approach for their numerical computation and to evaluate their performance by numerical simulations. The inherent questions on the asymptotic behaviour of these estimators (such as consistency, convergence rates or oracle inequalities) remain —to a large extent— unanswered. This opens an interesting area for future research.

A first attempt has been made in [21] where it is assumed that the model space $U \ni u^0$ is some Hilbert-space of real valued functions on some domain Ω and that $K : U \to L^2(\Omega)$ is linear and bounded. The error model (1) then has to be adapted accordingly. When Y is a Gaussian process on $L^2(\Omega)$ with mean Ku^0 and variance $\sigma^2 > 0$, consistency and convergence rates for SMREs as $\sigma \to 0^+$ have been proved in [21] for the case when $\Lambda = \mathrm{Id}$. However, in order to extend these results to the present setting, one would rather work with a discrete sample of Ku_0 on the grid X and then consider the case when the number of observations $N = md$ tends to infinity. The previous analysis in [21] indicates two major aspects that have to be considered in the asymptotic analysis for SMREs:

(a) As $N \to \infty$ usually the cardinality of the index set $\mathcal{S}$ (and hence of the set of weight functions $\mathcal{W}$) gets unbounded. Thus, the mutliresolution statistic $T(\varepsilon) = T_N(\varepsilon)$ in (4) is likely to degenerate unless it is properly normalized and $\mathcal{W}$ satisfies some *entropy condition*. In the linear case ($\Lambda = \mathrm{Id}$) we utilized a result from [17] that guarantees a.s. boundedness of $T_N(\varepsilon)$.

(b) In order to derive convergence rates (or risk bounds) it is well known that the true signal u^0 has to satisfy some apriori regularity conditions. When using general convex regularization functionals J, this is usually expressed by the *source condition*

$$K^* p^0 \in \partial J(u^0), \text{ for some } p^0 \in L^2(\Omega).$$

Here K^* denotes the adjoint of K and ∂J the (generalized) derivative of J. For example, if $J(u) = \frac{1}{2} \|u\|^2$, then this conditions means that $u^0 \in \mathrm{ran}(K^*)$.

It would be of great interest to transfer and extend the results in [21] to the present situation. It is to be expected that (a) and (a) above are necessary assumptions for this purpose.

As stressed by the referees, other extensions are of interest and will be postponed to future work. In contrast to imaging, in many other applications the design X is random, rather than fixed. In these situations an obvious way to extend our algorithmic framework would be to select suitable partitions $\mathcal{S}$ according to the design density, i.e. with finer resolution at locations with a high concentration of design points. It also remains an open issue how to extend the SMRE methodology to density estimation rather than regression, in particular in a deconvolution setup. For $d = 1$ a first step in this direction has been taken in [11] and it will be of great interest to explore whether our approach allow this to be extended to $d \geq 2$.

Acknowledgement

Appendix A. Proofs

In this section we shall give the proofs of Theorems 2.2 and 2.4 as well as Corollary 2.5. We note, that convergence of Algorithm 1 is a classical subject in optimization theory and a proof can e.g. be found in [20, Chap. III Thm. 4.1]. However, in order to apply these results, it is

necessary that certain regularity conditions for J hold, that are not realistic for our purposes (as e.g. in the case of total-variation regularization). The assertions of Theorems 2.2 and 2.4 are modifications of the standard results.

Moreover, we will allow for *approximate solution* of the subproblems (12) and (13). To this end, we rewrite these two subproblems as variational inequalities, i.e. given $(u_{k-1}, v_{k-1}, p_{k-1})$ we find (u_k, v_k, p_k) such that

$$G(v) - G(v_k) + \lambda^{-1} \langle Ku_{k-1} + v_k - Y - \lambda p_{k-1}, v - v_k \rangle \geq -\varepsilon_k, \ \forall v \in H \qquad (29a)$$

$$J(u) - J(u_k) + \lambda^{-1} \langle Ku_k + v_k - Y - \lambda p_{k-1}, Ku - Ku_k \rangle \geq -\delta_k, \ \forall u \in U \qquad (29b)$$

$$p_k = p_{k-1} - (Ku_k + v_k - Y)/\lambda, \qquad (29c)$$

where we assume that $\{\varepsilon_1, \varepsilon_2, \ldots\}$ and $\{\delta_1, \delta_2, \ldots\}$ are given sequences of positive numbers. Note that (29a) implies that $G(v_k) = 0$ and hence $v_k \in \mathcal{C}$ and that for $\varepsilon_k = \delta_k = 0$ (29a) and (29b) are equivalent to (12) and (13), respectively.

Finally, we remind the reader of the definition of the *subdifferential* (or generalized derivative) ∂F of a convex function $F : V \to \mathbb{R}$ on a real Hilbert-space V:

$$\xi \in \partial F(v) \quad \Leftrightarrow \quad F(w) \geq F(v) + \langle \xi, w - v \rangle_V \ \forall (w \in V).$$

If $\xi \in \partial F(v)$, then ξ is called *subgradient* of F at v. It follows from [19, Chap III, Prop. 3.1 and Prop. 4.1] that the Lagrangian L (and hence also the augmented Lagrangian L_λ) has a saddle-point $(\hat{u}, \hat{v}, \hat{p}) \in U \times H \times H$ if and only if

$$K\hat{u} + \hat{v} = Y, \quad K^*\hat{p} \in \partial J(\hat{u}) \quad \text{and} \quad \hat{p} \in \partial G(\hat{v}). \qquad (30)$$

We will henceforth assume that $\{(u_k, v_k, p_k)\}_{k \in \mathbb{N}}$ is a sequence generated by iteratively repeating the steps (29a) - (29c). Further, we introduce the notation

$$\bar{u}_k := u_k - \hat{u}, \quad \bar{v}_k := v_k - \hat{v} \quad \text{and} \quad \bar{p}_k := p_k - \hat{p}.$$

We start with the following

Lemma A.1. For all $k \geq 1$ we have that

$$\left(\|\bar{p}_{k-1}\|^2 + \lambda^{-2} \|K\bar{u}_{k-1}\|^2 \right) - \left(\|\bar{p}_k\|^2 + \lambda^{-2} \|K\bar{u}_k\|^2 \right)$$

$$\geq \lambda^{-2} \left(\|K\bar{u}_k + \bar{v}_k\|^2 + \|K\bar{u}_{k-1} - K\bar{u}_k\|^2 \right) - 2\lambda^{-2}(\delta_k + \delta_{k-1}) - \delta_k - \varepsilon_k \quad (31)$$

Proof. The assertion follows by repeating the steps (5.6)-(5.25) in the proof of [20, Chap. III Thm. 4.1] after replacing (5.9) and (5.10) by (29a) and (29b) respectively. □

We continue with the proof of Theorem 2.2. More precisely, we prove the following generalized version

Proposition A.2. *Assume that the sums $\sum_{k=1}^{\infty} \delta_k$ and $\sum_{k=1}^{\infty} \varepsilon_k$ are finite. Then, the sequence $\{(u_k, v_k)\}_{k \geq 1}$ is bounded in $U \times H$ and every weak cluster point is a solution of (8). Moreover,*

$$\sum_{k \in \mathbb{N}} \|Ku_k + v_k - Y\|^2 + \|K(u_k - u_{k-1})\|^2 < \infty.$$

Proof. Let $k \geq 1$ and define $D = \sum_{k=1}^{\infty} \delta_k$ and $E = \sum_{k=1}^{\infty} \varepsilon_k$. Summing up Inequality (31) over k and keeping in mind that $K\bar{u}_k + \bar{v}_k = Ku_k + v_k - Y$ and $K\bar{u}_{k-1} - K\bar{u}_k = Ku_{k-1} - Ku_k$ shows

$$\sum_{k=1}^{\infty} \|Ku_k + v_k - Y\|^2 + \|Ku_{k-1} - Ku_k\|^2$$

$$\leq \lambda^2 \|\hat{p}\|^2 + \|K\hat{u}\|^2 + (4\lambda^{-2} + 1)D + E < \infty$$

where we have used that $\bar{u}_0 = \hat{u}$ and $\bar{p}_0 = \hat{p}$. Furthermore, it follows again from (31) that

$$\|\bar{p}_k\|^2 + \lambda^{-2} \|K\bar{u}\|^2 \leq \|\hat{p}\|^2 + \lambda^{-2} \|K\hat{u}\|^2 + (4\lambda^{-2} + 1)D + E < \infty \tag{32}$$

This together with the fact that $\|Ku_k + v_k - Y\| \to 0$ shows that

$$\max(\|Ku_k\|, \|v_k\|, \|p_k\|) = \mathcal{O}(1).$$

Together with the optimality condition for (13) this in turn implies that for an arbitrary $u \in H$

$$J(u_k) \leq J(u) + \lambda^{-1} \langle Ku_k + v_k - Y - \lambda p_{k-1}, Ku - Ku_k \rangle + \delta_k = \mathcal{O}(1). \tag{33}$$

Summarizing, we find that

$$\max_{S \in \mathcal{S}} \mu_S(Ku_k - Y) + J(u_k) \leq \max_{S \in \mathcal{S}} \|\omega^S\| \|\Lambda(Ku_k - Y)\| + J(u_k) \leq c < \infty$$

for a suitably chosen constant $c \in \mathbb{R}$, since Λ is supposed to be continuous. Thus, it follows from Assumption A that $\{u_k\}_{k \in \mathbb{N}}$ is bounded and hence sequentially weakly compact. Now, let $(\tilde{u}, \tilde{v}, \tilde{p})$ be a weak cluster point of $\{(u_k, v_k, p_k)\}_{k \in \mathbb{N}}$ and recall that $(\hat{u}, \hat{v}, \hat{p})$ was assumed to be a saddle point of the augmented Lagrangian L_λ. Setting $u = \hat{u}$ in (33) thus results in

$$\begin{aligned} J(u_k) &\leq J(\hat{u}) + \lambda^{-1} \langle Ku_k + v_k - Y, K\hat{u} - Ku_k \rangle + \langle p_{k-1}, Ku_k - K\hat{u} \rangle + \delta_k \\ &= J(\hat{u}) + \langle p_{k-1}, Ku_k - K\hat{u} \rangle + \mathrm{o}(1) \end{aligned} \tag{34}$$

Using the relation $K\hat{u} + \hat{v} = Y$ we further find

$$\begin{aligned} \langle p_{k-1}, Ku_k - K\hat{u} \rangle &= \langle p_{k-1}, Ku_k - Y + \hat{v} \rangle \\ &= \langle p_{k-1}, Ku_k + v_k - Y \rangle - \langle p_{k-1}, v_k - \hat{v} \rangle = \mathrm{o}(1) - \langle p_{k-1}, v_k - \hat{v} \rangle \end{aligned} \tag{35}$$

From the definition of v_k in (29a) and from the fact that $\hat{v}, v_k \in \mathcal{C}$ it follows that

$$\langle Y + \lambda p_{k-1} - (Ku_{k-1} + v_k), \hat{v} - v_k \rangle \leq \bar{\varepsilon}_k$$

which in turn implies that

$$\begin{aligned} -\langle p_{k-1}, v_k - \hat{v} \rangle &\leq \lambda^{-1} \langle Y - (Ku_{k-1} + v_k), v_k - \hat{v} \rangle + \varepsilon_k \\ &= \lambda^{-1} \langle Y - (Ku_k + v_k), v_k - \hat{v} \rangle + \lambda^{-1} \langle Ku_k - Ku_{k-1}, v_k - \hat{v} \rangle + \varepsilon_k = \mathrm{o}(1) \end{aligned} \tag{36}$$

Combining (34), (35) and (36) gives

$$\limsup_{k \to \infty} J(u_k) \leq J(\hat{u}).$$

Now, choose a subsequence $\{u_{\rho(k)}\}_{k \in \mathbb{N}}$ such that $u_{\rho(k)} \rightharpoonup \tilde{u}$. Since J is convex and lower semi-continuous it is also weakly lower semi-continuous and hence the previous estimate yields

$$J(\tilde{u}) \leq \liminf_{k \to \infty} J(u_{\rho(k)}) \leq J(\hat{u}).$$

Moreover, we have that $v_{\rho(k)} \in \mathcal{C}$ for all $k \in \mathbb{N}$. Since $\mathcal{C}$ is closed we conclude that $\hat{v} \in \mathcal{C}$. Since $K\tilde{u} + \tilde{v} = Y$ this shows that $(\tilde{u}, \tilde{v})$ solves (8) and thus $J(\tilde{u}) = J(\hat{u})$. $\qquad\square$

We proceed with the proof of Theorem 2.4. Again we present a generalized version. To this end, let D and E be as in Proposition A.2.

Proposition A.3. *There exists a constant $C = C(\lambda, \hat{u}, \hat{v}, \hat{p}, E, D)$ such that*

$$0 \le J(u[\tau]) - J(\hat{u}) - \langle K^*\hat{p}, u[\tau] - \hat{u}\rangle_U \le C\tau + \delta_{k[\tau]} + \varepsilon_{k[\tau]} \quad \forall(\tau > 0).$$

Proof. Define $B^2 = (4\lambda^{-1} + 1)D + E$. Then it follows from (32) that

$$\lambda^{-1}\|Ku_k - K\hat{u}\| \le \|\hat{p}\| + \lambda^{-1}\|K\hat{u}\| + B \tag{37}$$

$$\|p_k\| \le 2\|\hat{p}\| + \lambda^{-1}\|K\hat{u}\| + B, \tag{38}$$

where $(\hat{u}, \hat{v}, \hat{p})$ is an arbitrary saddle point of $L_\lambda(u, v, p)$. Assume that $\tau > 0$ and that $k = k[\tau]$ is such that

$$\max(\|Ku_k + v_k - Y\|, \|Ku_{k-1} - Ku_k\|) \le \tau.$$

Then, it follows from (34), (36), (37) and (38) and that

$$\begin{aligned}
J(u_k) &\le J(\hat{u}) + \lambda^{-1}\langle Ku_k + v_k - Y, K\hat{u} - Ku_k\rangle + \langle p_{k-1}, Ku_k - K\hat{u}\rangle + \delta_k \\
&\le J(\hat{u}) + \tau(\|\hat{p}\| + \lambda^{-1}\|K\hat{u}\| + B) + \|p_{k-1}\|\tau + \langle p_{k-1}, \hat{v} - v_k\rangle + \delta_k \\
&\le J(\hat{u}) + \tau(3\|\hat{p}\| + 2\lambda^{-1}\|K\hat{u}\| + 2B) + 2\lambda^{-1}\tau\|v_k - \hat{v}\| + \delta_k + \varepsilon_k
\end{aligned} \tag{39}$$

After observing that $Ku_k - K\hat{u} + v_k - \hat{v} = Y$ it follows that $\|v_k - \hat{v}\| \le \tau + \|Ku_k - K\hat{u}\|$ and combining (39) and (37) gives

$$J(u_k) \le J(\hat{u}) + \tau\left(5\|\hat{p}\| + 4\lambda^{-1}\|K\hat{u}\| + 4B + 2\lambda^{-1}\tau\right) + \delta_k + \varepsilon_k.$$

Now, observe that from the definition of the subgradient and (30), it follows that $J(u_k) \ge J(\hat{u}) + \langle K^*\hat{p}, u_k - \hat{u}\rangle$ and that $\langle \hat{p}, v_k - \hat{v}\rangle \le 0$. This and the fact that $K\hat{u} + \hat{v} = Y$ implies that

$$\begin{aligned}
0 &\le J(u_k) - J(\hat{u}) - \langle K^*\hat{p}, u_k - \hat{u}\rangle \\
&= J(u_k) - J(\hat{u}) - \langle \hat{p}, Ku_k + v_k - Y\rangle + \langle \hat{p}, K\hat{u} + \hat{v} - Y\rangle + \langle \hat{p}, v_k - \hat{v}\rangle \\
&\le J(u_k) - J(\hat{u}) + \|\hat{p}\|\tau \\
&\le \tau\left(6\|\hat{p}\| + 4\lambda^{-1}\|K\hat{u}\| + 4B + 2\lambda^{-1}\tau\right) + \delta_k + \varepsilon_k.
\end{aligned} \tag{40}$$

This together with (39) finally proves the first part of the assertion. $\qquad\square$

Remark A.4. For the case when $D = E = 0$, it is seen from the proof of Proposition (A.3) that the constant C takes the simple form

$$C = \tau\left(6\|\hat{p}\| + \frac{4\|K\hat{u}\| + 2\tau}{\lambda}\right).$$

Proof of Corollary 2.5. Assume that $J(u) = \frac{1}{2}\|Lu\|_V^2$. Then it follows (see e.g. [22, Lem. 2.4]) that the subdifferential $\partial J(\hat{u})$ consists of the single element $L^*L\hat{u}$. Hence the extremality relations (30) imply that $K^*\hat{p} = L^*L\hat{u}$. Now it is easy to observe that

$$J(u_k) - J(\hat{u}) - \langle K^*\hat{p}, u_k - \hat{u}\rangle = \frac{1}{2}\|L(u_k - \hat{u})\|_V^2.$$

$\qquad\square$

References

[1] J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *J. Math. Imaging Vision*, 22(1):71–88, 2005.

[2] S. Becker, E. Cands, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3:165–218, 2011.

[3] M. Bertalmio, V. Caselles, B. Rougé, and A. Solé. TV based image restoration with local constraints. *J. Sci. Comput.*, 19(1-3):95–122, 2003. Special issue in honor of the sixtieth birthday of Stanley Osher.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[5] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in order restricted statistical inference (Iowa City, Iowa, 1985)*, volume 37 of *Lecture Notes in Statist.*, pages 28–47. Springer, Berlin, 1986.

[6] L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183, 2009.

[7] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[9] I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, 19(4):2032–2066, 1991.

[10] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29(1):1–65, 2001. With discussion and rejoinder by the authors.

[11] P. L. Davies and A. Kovac. Densities, spectral densities and modality. *Ann. Statist.*, 32(3):1093–1136, 2004.

[12] P. L. Davies, A. Kovac, and M. Meise. Nonparametric regression, confidence regions and regularization. *Ann. Statist.*, 37(5B):2597–2625, 2009.

[13] F. Deutsch and H. Hundal. The rate of convergence of Dykstra's cyclic projections algorithm: the polyhedral case. *Numer. Funct. Anal. Optim.*, 15(5-6):537–565, 1994.

[14] D. C. Dobson and C. R. Vogel. Convergence of an iterative method for total variation denoising. *SIAM J. Numer. Anal.*, 34(5):1779–1791, 1997.

[15] Y. Dong, M. Hintermüller, and M. Rincon-Camacho. Automated regularization parameter selection in multi-scale total variation models for image restoration. *J. Math. Imaging Vision*, 40:82–104(23), May 2011.

[16] L. Dümbgen and R. B. Johns. Confidence bands for isotonic median curves using sign tests. *J. Comput. Graph. Statist*, 13(2):519–533, 2004.

[17] L. Dümbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152, 2001.

[18] L. Dümbgen and G. Walther. Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785, 2008.

[19] I. Ekeland and R. Temam. *Convex analysis and variational problems*, volume 1 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-Oxford, 1976.

[20] M. Fortin and R. Glowinski. *Augmented Lagrangian methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983. Applications to the numerical solution of boundary value problems, Translated from the French by B. Hunt and D. C. Spicer.

[21] K. Frick, P. Marnitz, and A. Munk. Shape constrained regularisation by statistical multiresolution for inverse problems: Asymptotic analysis. Technical report, 2010. Available at *http://arxiv.org/abs/1003.3323*.

[22] K. Frick and O. Scherzer. Regularization of ill-posed linear equations by the non-stationary Augmented Lagrangian Method. *J. Integral Equations Appl.*, 22(2):217–257, 2010.

[23] N. Gaffke and R. Mathar. A cyclic projection algorithm via duality. *Metrika*, 36:29–54, 1989.

[24] M. Grasmair. The equivalence of the taut string algorithm and BV-regularization. *J. Math. Imaging Vision*, 27(1):59–66, 2007.

[25] D. M. Hawkins and R. Wixley. A note on the transformation of chi-squared variables to normality. *Amer.Statist.*, 40(4):296–298, 1986.

[26] S. W. Hell. Far-Field Optical Nanoscopy. *Science*, 316(5828):1153–1158, 2007.

[27] S. W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.*, 19(11):780–782, 1994.

[28] T. Hotz, P. Marnitz, R. Stichtenoth, L. Davies, Z. Kabluchko, and A. Munk. Locally adaptive image denoising by a statistical multiresolution criterion. *Comput. Stat. Data An.*, 56(3):543 – 558, 2012.

[29] G. M. James, P. Radchenko, and J. Lv. Dasso: connections between the dantzig selector and lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71:127–142, 2009.

[30] Z. Kabluchko. Extremes of the standardized Gaussian noise. *Stochastic Process. Appl.*, 121(3):515–533, 2011.

[31] Z. Kabluchko and A. Munk. Shao's theorem on the maximum of standardized random walk increments for multidimensional arrays. *ESAIM Probab. Stat.*, 13:409–416, 2009.

[32] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[33] Z. Lu, T. K. Pong, and Y. Zhang. An alternating direction method for finding Dantzig selectors. Technical report, 2010. Available at *http://arxiv.org/abs/1011.4604v1*.

[34] E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997.

[35] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations*, volume 22 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2001. The fifteenth Dean Jacqueline B. Lewis memorial lectures.

[36] T. Mildenberger. A geometric interpretation of the multiresolution criterion in nonparametric regression. *J. Nonparametr. Stat.*, 20(7):1048–5252, 2008.

[37] G. O. Mohler, A. L. Bertozzi, T. A. Goldstein, and S. J. Osher. Fast TV regularization for 2d maximum penalized likelihood estimation. *J. Comput. Graph. Statist*, 20(2):479–491, 2011.

[38] A. Munk, N. Bissantz, T. Wagner, and G. Freitag. On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):19–41, 2005.

[39] J. B. Pawley. *Handbook of Biological Confocal Microscopy*. Springer, 2006.

[40] J. Polzehl and V. G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(2):335–354, 2000.

[41] J. K. Romberg. The dantzig selector and generalized thresholding. In *CISS*, pages 22–25. IEEE, 2008.

[42] D. Siegmund and B. Yakir. Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213, 2000.

[43] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58:267–288, 1994.

[44] C. R. Vogel. *Computational methods for inverse problems*, volume 23 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. With a foreword by H. T. Banks.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[46] S. Xu. Estimation of the convergence rate of Dykstra's cyclic projections algorithm in polyhedral case. *Acta Math. Appl. Sinica (English Ser.)*, 16(2):217–220, 2000.